

**Genomic Context Analytics of Genes for Universal Stress Proteins from
Petroleum-Degrading *Alcanivorax***

Dissertation

**For the degree
MSc ENVIRONMENTAL SCIENCE**

**In the
DEPARTMENT OF ENVIRONMENTAL SCIENCES
University of South Africa**

by

**ZAINAB ABIMBOLA KASHIM
Student no: 55770975**

Prof. Raphael Isokpehi

Prof. Memory Tekere

September, 2016

TABLE OF CONTENTS

TABLE OF CONTENTS	i
LIST OF TABLES	iii
LIST OF FIGURES	iv
DECLARATION	vi
DEDICATION	vii
ACKNOWLEDGEMENT	viii
ABSTRACT	x
CHAPTER 1	1
INTRODUCTION	1
1.1 Overview of the genus <i>Alcanivorax</i>	1
1.2 Response to Stressor Factors.....	3
1.3 Databases on Functions of Universal Stress Proteins	4
1.4 Visual Analytics for Gaining Knowledge from Data on Universal Stress Proteins.....	5
1.5 Statement of the Problem	5
1.6 Justification	6
1.7 Hypothesis and Objectives.....	7
CHAPTER 2	8
LITERATURE REVIEW	8
2.1 Description of validly published species in the genus <i>Alcanivorax</i>	8
2.2 Genome Sequencing Projects of <i>Alcanivorax</i>	14
2.3 <i>Alcanivorax</i> Genes for Habitat-Related Stress Response	18
2.4 Universal Stress Protein Family	21
2.6 Bioinformatics Resources for Genomic Context Analytics	25
2.7 Visual Analytics and Bioinformatics.....	28
CHAPTER 3	32
MATERIALS AND METHODS	32
3.1 Objective 1: To make comparison among the protein domain organization of universal stress proteins encoded in <i>Alcanivorax</i> genomes.	33
3.1.1 Introduction	33
3.1.2 Retrieval of Datasets	33
3.1.3 Visual Analytics of Protein Domain Data.....	34
3.2 Objective 2: To evaluate the genomic context of genes encoding the universal stress protein	34
3.2.1 Introduction	34
3.2.2 Relative Transcription Direction of Genes for <i>Alcanivorax</i> Universal Stress Proteins	35
3.2.2 Chromosomal Cassette Search.....	35
3.3 Objective 3: To Predict biological processes involving <i>Alcanivorax</i> universal stress proteins.....	36
3.3.1 Evaluation of Stress Response Equipped Transcription Units.....	36
3.3.2 Gene Expression Patterns of Transcription Units Encoding <i>Alcanivorax</i> Universal Stress Proteins.....	37

CHAPTER 4.....	38
RESULTS	38
4.1 Objective 1: To make comparison among the protein domain organization of universal stress proteins encoded in <i>Alcanivorax</i> genomes.	38
4.1.1 Genome Statistics of <i>Alcanivorax</i> in the Integrated Microbial Genomes Database	38
4.1.2 Gene count and protein domain organization of universal stress proteins encoded in <i>Alcanivorax</i> genomes.	39
4.1.3 Protein sequence length and alignment of universal stress protein domain	41
4.2 Objective 2: To evaluate the genomic context of genes encoding the universal stress protein	45
4.2.1 Relative Transcription Direction of <i>Alcanivorax</i> Universal Stress Protein Genes	45
4.2.2 Chromosomal Cassette Search	46
4.3 Objective 3: To predict biological processes involving <i>Alcanivorax</i> universal stress proteins	49
CHAPTER 5.....	60
DISCUSSION AND CONCLUSIONS	60
5.1 Objective 1: To compare among the protein domain organization of universal stress proteins encoded in <i>Alcanivorax</i> genomes.	61
5.1.1 Genome Statistics of <i>Alcanivorax</i> in the Integrated Microbial Genomes Database	61
5.1.2 Gene count and protein domain organization of universal stress proteins encoded in <i>Alcanivorax</i> genomes.	62
5.2 Objective 2: To evaluate the genomic context of genes encoding the universal stress protein	64
5.2.1 Relative Transcription Direction of <i>Alcanivorax</i> Universal Stress Protein Genes	64
5.2.2 Chromosomal Cassette Search	65
5.3. Objective 3: To predict biological processes involving <i>Alcanivorax</i> universal stress proteins	66
5.3.1 Evaluation of Stress Response Equipped Transcription Units.....	66
5.3.3 Expression Levels of Genes in Transcription Units.....	68
5.3.4 Conclusions, Limitations of Findings and Future Work.....	68
REFERENCES	70

LIST OF TABLES

Table 1. Characteristics of members of the genus <i>Alcanivorax</i>	13
Table 2. <i>Alcanivorax</i> genomes sequencing: centers, methods and year.	15
Table 3. Isolation and ecosystem annotation of sequenced <i>Alcanivorax</i> genomes.	16
Table 4. Habitat-related stress response systems in <i>Alcanivorax borkumensis</i> SK2.....	19
Table 5. The list of genes for DNA repair and cell division of <i>Alcanivorax borkumensis</i> SK2...	20
Table 6. Findings from research project	38
Table 7. Genomes of <i>Alcanivorax</i> in the Integrated Microbial Genomes database.	39
Table 8. Count of genes in <i>Alcanivorax</i> genomes in the Integrated Microbial Genomes database.	40
Table 9. Chromosomal cassette alignment search results for selected <i>Alcanivorax</i> genes that encode universal stress proteins.....	48
Table 10. Functions of genes adjacent to genes for <i>Alcanivorax</i> universal stress proteins	49
Table 11. Sample information for GSE44687 gene expression profiling of <i>Alcanivorax</i> <i>borkumensis</i> cells: control vs. 1-octanol-stressed cells.....	55
Table 12. Probe identifiers for genes for universal stress proteins and adjacent genes in the Agilent-026176 <i>Alcanivorax borkumensis</i> SK2 microarray platform (GPL16725).....	56

LIST OF FIGURES

Figure 1. Neighbour-joining tree showing the phylogenetic positions of strain R8-12 ^T and type strains of some other related taxa, based on 16S rRNA gene sequences.....	2
Figure 2. Morphological features of selected <i>Alcanivorax</i> species.	8
Figure 3. Functional relatedness of <i>Alcanivorax</i> genomes based on Pfam annotation of genes. .	17
Figure 4. Chromosomal maps of genomes of <i>Alcanivorax borkumensis</i> SK2 and <i>Alcanivorax dieselolei</i> B5.....	18
Figure 5. A universal stress protein of <i>Porphyromonas gingivalis</i> is involved in stress responses and biofilm formation	20
Figure 6. Role of the 6 <i>E. coli</i> Usps in oxidative stress defense, iron metabolism, and cell surface properties.....	22
Figure 7. Deduced molecular structure of the lipopeptide produced by strain B-5.	24
Figure 8. Transmission electron micrograph of <i>Alcanivorax borkumensis</i> cells growing at a water – n-hexadecane interface.....	24
Figure 9. Overview of features of the Integrated Microbial Genomes (IMG) bioinformatics resource.....	25
Figure 10. BioCyc database collection of pathway/genome databases and software tools.	26
Figure 11. STRING (Search Tool for the Retrieval of Interacting Genes/Proteins)	27
Figure 12. Results page of search for functional interactions of a protein in STRING database.	28
Figure 13. Tableau Software user interface for conducting visual analytics.	30
Figure 14. Web-based resource for integrating annotation features for universal stress proteins of <i>Schistosoma mansoni</i> and <i>Schistosoma japonicum</i>	31
Figure 15. Overview of research approach for elucidating functions of universal stress proteins.	32
Figure 16. Pfam content of universal stress proteins in <i>Alcanivorax</i> genomes.	42
Figure 17. Distribution of amino acid sequence lengths of the <i>Alcanivorax</i> universal stress proteins.....	43
Figure 18. Protein domain organisation of <i>Alcanivorax</i> universal stress proteins.	44
Figure 19. Relative transcription direction of adjacent genes in <i>Alcanivorax</i> genomes.....	46
Figure 20. Gene adjacency profile and protein length of <i>Alcanivorax</i> universal stress proteins..	46
Figure 21. Example of chromosomal cassettes from non- <i>Alcanivorax</i> genomes with alignment to chromosomal cassette of <i>Alcanivorax borkumensis</i> SK2 containing universal stress protein ABO_1511.	47
Figure 22. Genomic context and transcription unit in <i>Alcanivorax borkumensis</i> SK2 universal stress protein ABO_1340 (ydaA).....	50
Figure 23. Genomic context and transcription unit in <i>Alcanivorax dieselolei</i> B5 universal stress protein B5T_02274 (ydaA).....	50
Figure 24. Multi-genome alignment of neighborhood of genes for universal stress protein of <i>Alcanivorax dieselolei</i> and selected genomes.....	51

Figure 25. Interaction views from <i>Alcanivorax borkumensis</i> locus tags ABO_1340, ABO_2141, ABO_1511 and ABO_1011	52
Figure 26. Confidence scores for interactions with functional partners to <i>Alcanivorax borkumensis</i> universal stress protein	53
Figure 27. Average expression levels of <i>Alcanivorax borkumensis</i> genes for universal stress protein and adjacent genes.	57
Figure 28. Overview of expression levels of <i>Alcanivorax borkumensis</i> probes for genes annotated for universal stress protein domain and their adjacent genes in same transcription direction.	58
Figure 29. Interface for discovering expression levels of probes of <i>Alcanivorax borkumensis</i> for universal stress protein and adjacent genes.	59

DECLARATION

I Zainab Abimbola Kashim hereby declare that the dissertation/thesis, which I hereby submit for the degree of Masters In Environmental Science at the University of South Africa, is my own work and has not previously been submitted by me for a degree at this or any other institution.


I declare that the dissertation /thesis does not contain any written work presented by other persons whether written, pictures, graphs or data or any other information without acknowledging the source.

I declare that where words from a written source have been used the words have been paraphrased and referenced and where exact words from a source have been used the words have been placed inside quotation marks and referenced.

I declare that I have not copied and pasted any information from the Internet, without specifically acknowledging the source and have inserted appropriate references to these sources in the reference section of the dissertation or thesis.

I declare that during my study I adhered to the Research Ethics Policy of the University of South Africa, received ethics approval for the duration of my study prior to the commencement of data gathering, and have not acted outside the approval conditions.

I declare that the content of my dissertation/thesis has been submitted through an electronic plagiarism detection program before the final submission for examination.

Student signature:  _____

Date: _____

DEDICATION

This thesis is dedicated to God Almighty Allah for making this work a successful one.

To my dear parents Alhaji and Hajiya Kashim, my lovely husband Dr Abdulhakeem Bello and Son Abdurrahman Abidemi Bello whose words of encouragement, prayers and support kept me going all the way.

ACKNOWLEDGEMENT

All Glory and Honour to Almighty Allah who gave me the opportunity to start this important journey, and saw me through till the end. Unto Thee alone do I give my profound gratitude for being alive.

I wish to express my unreserved gratitude to my team of supervisors, Prof. R. Isokpehi and Prof. M. Tekere for their advice and guidance in ensuring the completion and success of this programme.

My sincere appreciation goes to my wonderful parents, Alhaji and Hajiya Kashim for their inspirational words, financial and moral support. To my siblings and family members who offered invaluable assistance and support, May Allah reward you. My deep appreciation and love goes to my dearest husband Dr Abdulhakeem Bello and son Abdurrahman Abidemi Bello for standing by me all through the work, I say thanks for always being there. My lovely in-laws and the Bello's dynasty at large, I appreciate your prayers and support.

I want to acknowledge the Director of Genetics Genomics and Bioinformatics NABDA Prof. Oyekanmi Nash for his fatherly advice and support throughout the work, colleagues and well-wishers, I appreciate you all.

ABSTRACT

Alcanivorax species are gram negative bacteria that usually require aliphatic hydrocarbon as the sole carbon source for growth. The ability to use petroleum in polluted environments as energy source makes *Alcanivorax* species biotechnologically relevant in bioremediation. Universal stress proteins confer ability to respond to unfavourable environments, thus the present study was done to analyse the genomic context of genes for universal stress proteins in *Alcanivorax* genomes. A combination of bioinformatics and visual analytics approaches were used to analyze genome-enabled data including sequences and gene expression datasets. On the basis of transcription unit and adjacent genes, two types of *Alcanivorax* USP genes observed were (i) adjacent to cyclic nucleotide-binding and oxygen sensing functions; and (ii) adjacent to sulfate transporter function. Both types of genes encode two universal stress protein domains (pfam00582) also referred to as tandem-type universal stress proteins. The sequence and structural characteristics of each of the four USP domains in *Alcanivorax* needs to be further investigated. This dissertation research evaluated data from *Alcanivorax borkumensis* cells (grown on either pyruvate or hexadecane as carbon source) that were stressed with 1-octanol and data collected at 15 min, 30 min, 60 min and 90 min after 1-octanol addition. The two genes for *Alcanivorax borkumensis* SK2 universal stress proteins, ABO_1340 and ABO_1511, had the same direction of expression for adjacent genes. A limitation of this research was that findings based on bioinformatics and visual analytics methods may need confirmation by molecular methods. The differences observed may also reflect the quality of the annotations provided for genes. The sequence and structural characteristics of each of the four USP domains in *Alcanivorax* needs to be further investigated. Further research is needed on the relationship between number, length and order of genes in operons that include genes for universal stress

proteins. Additionally, *in vitro* studies to confirm the functional prediction made from the genomic context of the universal stress protein in *Alcanivorax* genome. The knowledge discovered from this genome context analytics research could contribute to improving the performance of *Alcanivorax* species in bioremediation of environments polluted with petroleum.

CHAPTER 1

INTRODUCTION

1.1 Overview of the genus *Alcanivorax*

Members of the Gram-negative bacteria genus *Alcanivorax* use aliphatic hydrocarbon as sole or principal carbon sources for growth (Fernández-Martínez *et al.*, 2003; Yakimov *et al.*, 1998). Species are obligate aerobes, either non-motile and non-flagellated or motile by polar flagella and halophilic, requiring (at least) Na⁺ ions for growth: some species have more complex ionic requirements (Fernández-Martínez *et al.*, 2003).

The type species of the genus is *Alcanivorax borkumensis* which was isolated from the island of Borkum, North Sea, located close to the German-Dutch border (Yakimov *et al.*, 1998). The taxonomic classification of *Alcanivorax borkumensis* is as follows. *A. borkumensis* belongs to the Kingdom Bacteria, Phylum Proteobacteria, Class Gamma Proteobacteria, Order *Oceanospirillales*, Family *Alcanivoracaceae*, and Genus *Alcanivorax*. The etymology of the genus name is N.L. masc. n. *alcanum*, alkane, aliphatic hydrocarbon; L. adj. *vorax*, devouring, ravenous, voracious; N.L. masc. n. *Alcanivorax*, alkane-devouring (Parte, 2013). As of August 2014, the nine validly published species of *Alcanivorax* were *A. balearicus* (Rivas *et al.*, 2007), *A. borkumensis* (Yakimov *et al.*, 1998), *A. dieselolei* (Liu and Shao, 2005), *A. hongdengensis* (Wu, *et al.*, 2009), *A. jadensis* (Fernández-Martínez *et al.*, 2003), *A. marinus* (Lai *et al.*, 2013), *A. pacificus* (Lai *et al.*, 2011a), *A. venustensis* (Fernández-Martínez *et al.*, 2003) and *A. xenomutans* (Rahul *et al.*, 2014). A phylogenetic relationship within the genus *Alcanivorax* and related taxa is presented in Figure 1 (Lai *et al.*, 2013).

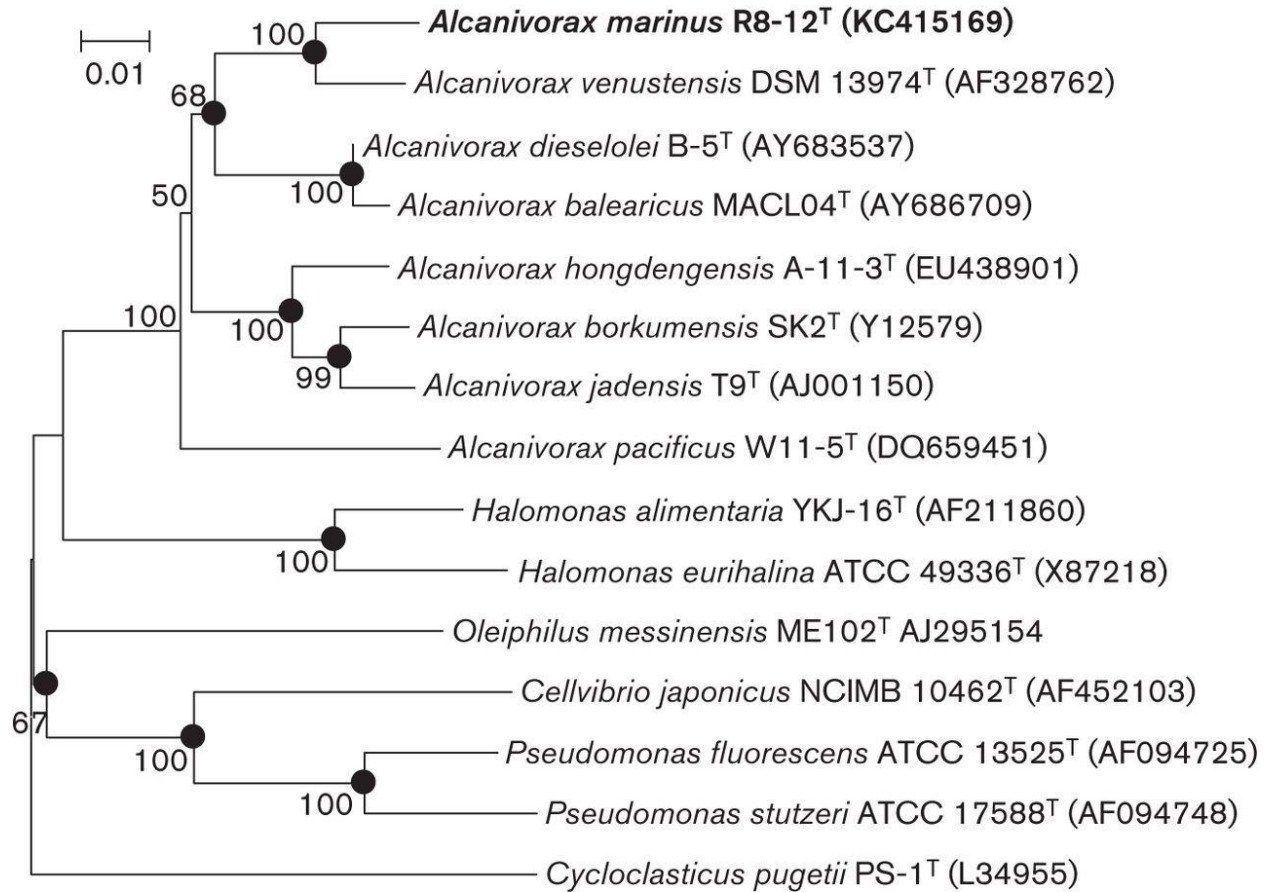


Figure 1. Neighbour-joining tree showing the phylogenetic positions of strain R8-12^T and type strains of some other related taxa, based on 16S rRNA gene sequences.

This figure is based on the description of *Alcanivorax marinus*. Filled circles indicate nodes that were also recovered in maximum-likelihood and minimum-evolution trees based on the same DNA sequences. Bootstrap values (expressed as percentages of 1000 replications) are shown at branch points. Bar, 0.01 indicates the nucleotide substitution rate (K_{nuc}) units. Source: (Lai et al., 2013).

Alcanivorax borkumensis is a Gram-negative, rod-shaped, non-motile, aerobic and halophilic bacterium that use alkanes, major constituents of crude oil, as the primary source of energy to produce glucose lipids biosurfactants (Kasai *et al.*, 2002; Yakimov *et al.*, 1998). It is found predominantly in sea water and oceanic environment which contain petroleum oil that could potentially result from oil spill or other sources. It can also be found in minute amount of non-polluted water (Wang and Shao, 2013). This unique species can flourish in the presence of large solid matter such as sand or gravel (Kasai *et al.*, 2002). The habitat of

Alcanivorax borkumensis is usually on or near the surface of water where it utilizes the compound in the oil as source of energy and can withstand salinity of about 1% to 12.5% and temperatures of 4 °C to 35 °C, with optimum temperature between 25 °C and 30 °C (Yakimov *et al.*, 1998). The distinctive features of *A. borkumensis* from other species might be the flexibility of its DNA and metabolism, which allows the bacteria to outcompete other bacteria species, and this is because it consumes a wider variety of alkanes than other known species (Hara *et al.*, 2003).

1.2 Response to Stressor Factors

A unique feature of *A. borkumensis* is the ability to withstand harsh environments such as its halotolerant capability which works on the principle of accumulating potassium ion, glutamate and the solutes ectoine and betaine as osmoprotectants (Sabirova *et al.*, 2008). It also has the ability to protect itself against oxidative stress resulting from UV radiation, and to survive under low temperatures (Sabirova *et al.*, 2008).

The technologies for high-throughput determination of genomic sequences and protein abundance is providing data on genome-encoded pathways, genomic features and encoded pathways for diverse applications including surveillance and detection of microorganisms (Liu *et al.*, 2009; Nakamura *et al.*, 2009). The volumes of biological sequence data led researchers in bioinformatics to develop the Gene Ontology (GO) as an approach for analysing and sharing the functional annotation of genes and proteins (Ashburner *et al.*, 2000). The GO is a dynamic controlled vocabulary of over 16, 000 terms used to describe molecular function, process and location of activity of a protein in a generic cell (Consortium, 2004). The term “Response to Stress” has been mapped to protein domains in the Protein Family (Pfam) Database (Finn *et al.*, 2013). One of the gene families annotated for response to stress is the genes that encode the

universal stress protein domain (accession number PF00582 or pfam00582). The mappings are available at <http://geneontology.org/external2go/pfam2go>.

Furthermore, the genes encoding the universal stress protein (USP) domain have been detected in diverse organisms, including plants, bacteria and fungi (Hingley-Wilson *et al.*, 2010, Kvint, *et al.*, 2003). Known functions of universal stress proteins include promoting organismal survival during extreme conditions such as drought (Isokpehi *et al.*, 2011b), osmolarity (Schweikhard *et al.*, 2010), oxidative stress due to hydrogen peroxide (Isokpehi, *et al.*, 2011) and acid response (Gury *et al.*, 2009). The universal stress proteins have been classified broadly based on their ability to bind or not bind Adenosine Tri-Phosphate (ATP) (Sousa and McKay, 2001, Zarembinski *et al.*, 1998).

1.3 Databases on Functions of Universal Stress Proteins

The assignment of functions to protein-coding genes from genome sequences is a key process to understanding encoded biological systems of an organism. A U.S. National Research Council Report on Sequence-Based Classification of Select Agents identified that, a goal for advancing the understanding of biological systems is the ability to predict accurately the function of individual proteins from genome sequence, including what ligands or macromolecules they bind to (National Research Council, 2010). Protein sequences can consist of one or more domains, which are groups of amino acids that represent structural, functional and evolutionary units of the protein (Wang and Caetano-Anollés, 2009). Information on protein sequences that can be used to understand their functions are stored in biological databases.

More than 13,000 protein families are available in the Pfam Protein Family database and these protein domains have been utilized to assign functions to proteins (Finn *et al.*, 2013). The

Conserved Domain Database (CDD) at the National Centre for Biotechnology Information (NCBI) has predictions of functional sites including biologically relevant chemical ligand binding sites of individual proteins (Marchler-Bauer *et al.*, 2011). The Integrated Microbial Genomes (IMG) System serves as a community resource for comparative analysis and annotation of all publicly available genomes from three domains of life in a uniquely integrated context. The IMG system can be used to access genome sequences and gene context (Markowitz *et al.*, 2012, Mavromatis *et al.*, 2009).

1.4 Visual Analytics for Gaining Knowledge from Data on Universal Stress Proteins

Visual analytics is a means of exploring visually and understanding data of any size (Hanrahan *et al.*, 2009). Visual analytics is described as an interactive process conducted via visual interfaces that involves collecting information, data pre-processing, knowledge representation, interaction, and decision-making (Johnson *et al.*, 2010, Thomas and Cook, 2006). The goals of visual analytics include the discovery of new patterns, hypothesis generation, confirmation of assumed patterns, and promotion of quantitative reasoning (Carr and Pickle, 2011). The increasing availability of large datasets from microbial genomics for disease outbreak surveillance provides new opportunities to use visual analytics (Sims *et al.*, 2011).

1.5 Statement of the Problem

In the marine environment, crude oil is subjected to physico-chemical and biological modifications, which aids hydrocarbons solubility in the water and consequentially cause extensive damage to marine life, natural resources, and human health. For example, it is estimated that most petroleum compounds have carcinogenic properties (Aas *et al.*, 2000, Ericson *et al.*, 1998). The toxicity of several hydrocarbons has led to their classification by the Agency of Environmental Protection, the World Health Organization, and the European Union as

top priority pollutants. Furthermore, due to the geographic distribution and the high toxicity of hydrocarbons, they are considered as the principal organic markers of the anthropogenic activity in the ecosystems (Peterson *et al.*, 1996, Zhang *et al.*, 2014). Spilled crude-oil which is denser than water reduces and restricts its permeability: organic hydrocarbons which fill the soil pores expel water and air, thus depriving the plant roots of the much needed water and air (Nicolotti and Egli, 1998; Nwankwo *et al.*, 2014).

The removal of petroleum hydrocarbon or its transformation into less toxic products by bioremediation is a less invasive and less expensive process if compared to classical decontamination. However, the use and optimization of bioremediation treatments in the water compartment require knowledge of the seawater microbial communities directly and indirectly involved in the degradation of hydrocarbons (Harayama *et al.*, 2004). The degradation process can be time-consuming to achieve when depending on the natural inhabitants of the water and oil spillage to degrade the oil. The amount of data generated from genes, proteins and molecular mechanisms for degradation and response to the toxic environment is ever growing (Mason *et al.*, 2014, Sierra-García *et al.*, 2014).

1.6 Justification

The genomic context of the genes for universal stress proteins of *Alcanivorax* species can help reveal the mechanisms of resistance of *Alcanivorax* species to the atypical conditions in oil polluted environment. The insights could enhance biotechnological research on oil degrading bacteria. The comprehension of large datasets from microbial genomics and metagenomics research will be enhanced with the use of the interactive platform provided by visual analytics for knowledge-building, sense-making and decision-making.

Universal stress protein genes that are predicted to encode stress tolerance in genomes of *Alcanivorax* will be identified using bioinformatics methods and characterized with respect to abundance, sequence homology, functional protein domain diversity, predictive biochemical properties and chromosomal context.

1.7 Hypothesis and Objectives

The hypothesis of the research is that the integrated application of bioinformatics and visual analytics methods will help uncover the biological processes involving the universal stress proteins of *Alcanivorax* bacteria. The objectives were to

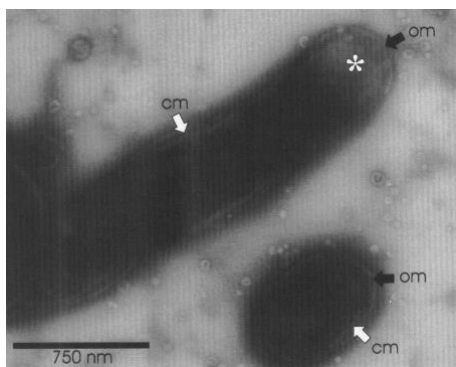
1. Make comparison among the protein domain organization of universal stress proteins encoded in *Alcanivorax* genomes.
2. Evaluate the genomic context of genes encoding the universal stress protein
3. Predict biological processes involving *Alcanivorax* universal stress proteins.

CHAPTER 2

LITERATURE REVIEW

2.1 Description of validly published species in the genus *Alcanivorax*

There are publications describing the nine known *Alcanivorax* species which provide details of genotypic and phenotypic characteristics of the genus (Fernández-Martínez *et al.*, 2003, Lai *et al.*, (2011a), Lai *et al.*, (2013b), Liu and Shao, 2005, Parte 2014; Rahul *et al.*, 2014, Rivas *et al.*, 2007, Wu *et al.*, 2009, Yakimov *et al.*, 1998). Some aspects of the descriptions are extracted for each species in the following paragraphs. Shared abilities of the species include (i) the ability to use hydrocarbons as sole carbon source; (ii) ability to produce extracellular lipids in conditions of carbon excess and limited nitrogen availability such as oil spills and (iii) ability to grow in the saline environments. The morphology of a non-motile and motile species of *Alcanivorax* is shown in Figure 2.



a. *Alcanivorax borkumensis* SK2 cells



b. *Alcanivorax jadensis* T9.

Figure 2. Morphological features of selected *Alcanivorax* species.

a. Negatively stained mid-exponential-phase *Alcanivorax borkumensis* SK2 cells appear as rods. Because of osmotic sensitivity the periplasmic space at the apex of the cells has dilated and appears electron-translucent (asterisk). Outer membrane (om) and cytoplasmic membrane (cm) are indicated (Yakimov *et al.*, 1998). b. Transmission electron micrographs of a cell of *Alcanivorax jadensis* T9 from a 4 d culture showing pili; bar, 0.39 (Bruns and Berthe-Corti, 1999).

Alcanivorax cells are well known and have properties such as a rod shape morphology, generally wide (0.6–0.8 µm) and long (1.6–2.5 µm) when growing in pyruvate-supplemented medium but short (1.0–1.5 µm) when n-alkanes are used as single carbon and energy source (Yakimov *et al.*, 1998). *Alcanivorax borkumensis* SK2 is the first and most extensively studied member of this genus and harbours two alkB genes and three P450 genes (Schneiker *et al.*, 2006). It was isolated from a seawater sediment in the North Sea at a site located near the Isle of Borkum; species are Gram-negative, aerobic, rods (1.6–2.5×0.6–0.8 µm), feed on alkane as source of organic compounds and energy (Schneiker *et al.*, 2006). Halophilic species are able to grow in the presence of 12% NaCl. The principal fatty acids presents are C16: 0, C18: 1 ν 7c and C16: 1 ν 7c. The DNA G+C content of species is 54.7mol%, and on the basis of 16S-rDNA-based phylogenetic analyses, the genus belongs to the γ -subclass.

Alcanivorax hongdengensis strain A-11-3^T was isolated from an oil-enriched consortium from the surface seawater of Hong-Deng dock in the Straits of Malacca (Wu *et al.*, 2009). They are aerobic, Gram-negative, non-spore forming with irregular rod shape (0.9–3.2×0.5–0.8 µm) and catalase and oxidase negative. *A. hongdengensis* usually grow on a restricted spectrum of organic compounds which may include organic acids and alkanes. Comparison of 16S rRNA gene sequence indicated that strain the A-11-3^T was mostly related to strains of *Alcanivorax jadensis* (96.8% sequence similarity), *Alcanivorax borkumensis* (96.8%), and *Alcanivorax balearicus* (94.0%). The most common fatty acids present are C_{16:0} (31.2%), C_{18:1 ω 7c} (24.8%), C_{18:0} (9.6%), C_{12:0} (8.3%), C_{16:1 ω 7c} (8.3%) and C_{16:0}3-OH (5.1%). The G+C composition of the genomic DNA is 54.7 mol% produces lipopeptides and utilizes alkanes ranging from C8 to C36. The ability of *A. hongdengensis* A-11-3 to survive in oil-polluted environments could be due to its multiple systems for alkane degradation and its range of substrates.

Alcanivorax pacificus also called strain W11-5^T was described by Lai *et al.*, 2011a, isolated from a pyrene-degrading consortium enriched deep-sea sediment of the Pacific Ocean (Lai *et al.*, 2011a). The organism possess properties which include, rod-shaped (1.7–2.3×0.3 μm), negative stain, non-motile and positive to both oxidase and catalase test. It is slightly halophilic and grows between 10–42 °C with optimum growth in the range of 25–28 °C. Based on the 16S rRNA gene sequence analysis, the strain W11-5^T was shown to belong to the genus *Alcanivorax* closely related to *A. dieselolei* B-5^T (93.9% 16S rRNA sequence similarity), *A. balearicus* MACL04^T (93.1%), *A. hongdengensis* A113^T (93.1%), *A. borkumensis* SK2^T (93.0%), *A. venustensis* ISO4^T (93.0%) and *A. jadensis* T9^T (92.9%). The major fatty acids present are C_{12:0} 3-OH (8.0%), C_{16:0} (29.1%) and C_{18:1}ω7c (27.4%), with G+C content of 60.8 mol%.

Alcanivorax balearicus also known as strain type MACL04 was isolated in 2007 by Rivas *et al.*, from Lake Martel, a subterranean saline lake in Mallorca (Spain) (Rivas *et al.*, 2007). The organism has a rod-shape (1.1–1.8×0.6–0.8 μm), uses flagellum for movement, appears opaque and mucoid with blue-green iridescence when grown in 1% v/v Tween 20. The major fatty acids contents are C_{19:0} cyclo ω8c and C_{16:0}. The sequences of the large and short 16S–23S intergenic spacer regions showed similarities of 97.2 and 98.8% (ungapped) with respect to *A. dieselolei* B-5^T. Partial sequences of *gyrB* and *alkB* genes showed 94.0% similarity between strain MACL04^T and *A. dieselolei* B-5^T. The G+C content of strain MACL04^T was 62.8 mol%.

Alcanivorax dieselolei strain B-5^T and N01A isolated from the surface water of Bohai and deep-sea sediment of the East Pacific Ocean respectively. The organisms had the following properties: rod-shaped (0.9–1.8×0.3–0.5 μm), Gram-negative, halophilic, motile, aerobic, and

positive to oxidase and catalase tests. From the analysis of 16S rRNA gene sequence, strains B-5^T and NO1A have been demonstrated to belong to the *γ-Proteobacteria*, with similarities found with *Alcanivorax venustensis* (95.2%), *Alcanivorax jadensis* (94.6%) and *Alcanivorax borkumensis* (94.1%). Principal fatty acids presents are C_{16:0}, C_{16:1ω7c} and C_{18:1ω7c}. Chemotaxonomically characteristic fatty acid C_{19:0} cyclo ω8c was also detected. The type strain of *Alcanivorax dieselolei* is B-5^T (=DSM 16502^T=CGMCC 1.3690^T). Recently, (Liu and Shao, 2005) showed that in the type strain *A. dieselolei* B-5, four genes, including two *alkB*, one *P450* and one *almA*, are involved in alkane assimilation. Thus, *A. dieselolei* is a promising candidate for oil pollution mitigation due to its wide substrate range and hydrocarbon-degrading abilities (Liu and Shao, 2005, Lai *et al.*, 2011).

Alcanivorax marinus strain R8-12^T is a Gram-negative rod-shaped (1.0–1.2×0.5–0.6 μm), positive to catalase and oxidase test which can survive in temperatures range of 10-42 °C with optimum growth at 28 °C and pH values ranging from 6-10 (Lai *et al.*, 2013). Based on the 16S rRNA gene sequence analysis, strain R8-12^T belongs to the genus *Alcanivorax* and closely related to *Alcanivorax venustensis* DSM 13974^T (97.2%), *A. dieselolei* B-5^T (95.0%), *A. balearicus* MACL04^T (94.6%), *A. hongdengensis* A-11-3^T (94.3%), *A. jadensis* T9^T (93.8%), *A. borkumensis* SK2^T (93.7%) and *A. pacificus* W11-5^T (93.7%). The *gyrB* sequence similarities between R8-12^T and other species of the genus *Alcanivorax* ranged from 77.9% to 86.9%. The major fatty acids present are C_{16:0} (31.8%), C_{18:1ω7c} (20.3%), C_{19:0}ω8c cyclo (15.8%) and summed feature 3 (C_{16:1ω6c} and/or C_{16:1ω7c}) (8.9%).

Alcanivorax venustensis was isolated from ‘Portus Venustus’ Elegant Port. The organism had the following properties: rod shaped (0.9-1.8×0.3-0.5 μm), aerobic, Gram-negative, motility is by flagella and non-pigmented colonies (Fernández-Martínez *et al.*, 2003). It grows within a

temperature range of 4-40 °C and requires marine salt for growth with G+C content of 66mol%. They major fatty acids present includes: C16 : 0, C16 : 1w7c, C18 : 1w7c and C19 : 0cyclo, with minor traces of 3OH-C12 : 0, C12 : 0 and C10 : 0. The closely related strain to *A. venustensis* based on sequence analysis (16SrDNA) is *A. borkumensis* and *Alcanivorax jadensis* T9^T previously known as *Fundibacter jadensis* (Yakimov *et al.* 1998, Bruns and Berthe-Corti 1999).

Alcanivorax jadensis T9^T is a slightly halophilic hydrocarbon-degrading bacterium isolated from continuous cultures having a suspension of intertidal sediment from the German North Sea coast with hexadecane as the main carbon source (Bruns and Berthe-Corti, 1999). It is a Gram-negative, aerobic, rod-shaped (0.8–1.8×0.3–0.7 μm) that grows at concentrations of 0.5-15% (w/v) NaCl and utilizes a restricted spectrum of carbon sources. The G+C content of the DNA is 63.6mol%.

Alcanivorax xenomutans is the most recently published species of the *Alcanivorax* genus (Rahul *et al.*, 2014). The cells possess the wee known properties which are as follows: they are motile rods, positive to catalase and oxidase test, Gram- negative, hydrolyze Tween 80, grow chemoorganoheterotrophically with an optimal pH of 6 (range 4-9) and at 30 °C (range 25-40 °C). Analysis based on 16S rRNA demonstrated similarities between them and the *Alcanivorax* genus, which are the closest neighbors to *Alcanivorax dieselolei* B-5T (sequence similarity values of 99.3 and 99.7%, respectively) and *Alcanivorax balearicus* MACL04T (sequence similarity values of 98.8 and 99.2%, respectively). A comparison of characteristics of selected *Alcanivorax* species is presented in Table 2. Six of the *Alcanivorax* species are unable to ferment d-glucose. The *A. pacificus* strain possess is only species that possess the β-Glucosidase enzyme.

Table 1. Characteristics of members of the genus *Alcanivorax*

Characteristic	Members of genus <i>Alcanivorax</i> ^a						
	1	2	3	4	5	6	7
Catalase	+	+	+	-	+	+	-
Oxidase	+	W	W	-	+	W	+
Motility, flagella arrangement	-	+*	+†	-	+*	-	-
Ionic requirements	Na+	Na+	-	Na+	Complex	Complex	Na+
Growth in 17% NaCl	-	W	-	-	+	-	-
Growth at 42 °C	+	+	-	+	+	-	+
Growth at 45 °C	-	+	-	-	+	-	+
DNA G+C content (mol%)	60.8	62.1	62.8	54.7	66.4	53.4	63.6
API 20 NE:							
Nitrate reduction	+	-	W	+	-	-	-
d-Glucose fermentation	W	-	-	-	-	-	-
Urease, β-glucosidase, d-glucose, l-arabinose, N-acetylglucosamine	+	-	-	-	-	-	-
Gelatin hydrolysis	-	-	-	+	-	+	-
Capric acid, phenylacetic acid	-	+	+	-	-	-	-
Adipic acid	W	+	+	-	+	-	-
Malic acid	+	+	-	-	-	-	-
Trisodium citrate	-	+	+	-	-	-	-
API ZYM:							
Acid phosphatase, naphthol-AS-B1-phosphoamidase	W	+	+	w	+	+	+
Alkaline phosphatase	W	+	+	-	w	+	+
β-Glucosidase	+	-	-	-	-	-	-

^a Taxa: 1, *Alcanivorax pacificus* W11-5T (Lai *et al.*, 2011); 2, *A. dieselolei* B-5T [data from (Liu and Shao, 2005)]; 3, *A. balearicus* MACL04T (Rivas *et al.*, 2007); 4, *A. hongdengensis* A-11-3T (Wu, Lai, Zhou, Qiao, & Liu, 2009); 5, *A. venustensis* ISO4T (Fernández-Martínez *et al.*, 2003); 6, *A. borkumensis* SK2T (Yakimov *et al.*, 1998); 7, *A. jadensis* T9T (Bruns and Berthe-Corti, 1999, Fernández-Martínez *et al.*, 2003). Tests for catalase and oxidase activities and tests in the API 20 NE and API ZYM systems were performed in parallel for all seven type strains. In the API 20 NE system, all strains were negative for denitrification, indole production, arginine dihydrolase and β-galactosidase activities and for the utilization of D-mannose, D-mannitol, maltose and potassium gluconate. In the API ZYM system, all strains were positive for esterase (C4), esterase lipase (C8), lipase (C14), leucine aminopeptidase, weakly positive for valine aminopeptidase and negative for cystine aminopeptidase, N-acetyl-β-glucosaminidase, trypsin, α-chymotrypsin, α-fucosidase, α- and β-galactosidase, α-glucosidase, α-mannosidase and β-glucuronidase. +, Positive; w, weakly positive; -, negative. *Lophotrichous flagella; †Polar or subpolar flagellum

Source: (Lai *et al.*, 2011)

2.2 Genome Sequencing Projects of *Alcanivorax*

A purpose of genome sequencing projects is to provide, for a genome, nucleotide sequence data and associated annotations including genes and their functions (Markowitz, Chen, Palaniappan, *et al.*, 2012). Such data present opportunities for comparative analysis of genomes of closely related organisms, for example the genomes obtained from species in the same bacteria genus (Simmons *et al.*, 2011). The advent of next generation sequencing platforms including 454 GS FLX Titanium and Illumina MiSeq has revolutionized the speed of genome sequencing and availability of genome sequences in public domain databases (Kyrpides *et al.*, 2014).

For example, there are 10 genome sequences of *Alcanivorax* species available in the Integrated Microbial Genomes (IMG) system (Table 3) (Markowitz *et al.*, 2014). As of August 2014, *Alcanivorax dieselolei* and *Alcanivorax borkumensis* were the two finished genome available. Other validly published *Alcanivorax* with genome sequences are *A. hongdengensis* and *A. pacificus*. The first published genome sequence of an *Alcanivorax* species was that of the type species *Alcanivorax borkumensis* SK2 (Schneiker *et al.*, 2006). In August 2014, the complete genome of *Alcanivorax sp.* strain NBRC 101098 was published (Miura *et al.*, 2014).

Table 2. *Alcanivorax* genomes sequencing: centers, methods and year.

Genome	Sequencing Center	Sequencing Method	Year of Sequencing
<i>Alcanivorax borkumensis</i> SK2	Bielefeld University	Sanger	2006
<i>Alcanivorax dieselolei</i> B5	Third Institute of Oceanography, State Oceanic Administration, China	454	2013
<i>Alcanivorax hongdengensis</i> A-11-3	The Third Institute of State Oceanic Administration (SOA) Third Institute of Oceanography, State Oceanic Administration	Illumina GA IIX	2013
<i>Alcanivorax pacificus</i> W11-5	Functional Genomics Center, Zurich	Illumina GA IIX	2013
<i>Alcanivorax sp.</i>			2011
<i>Alcanivorax sp.</i> 43B_GOM- 46m	DOE Joint Genome Institute First Institute of Oceanography, State Oceanic Administration of China	Illumina HiSeq 2000, Illumina HiSeq 2500	2013
<i>Alcanivorax sp.</i> 97CO-5		Illumina HiSeq	2014
<i>Alcanivorax sp.</i> DG881	J. Craig Venter Institute		2010
<i>Alcanivorax sp.</i> JRC	Univ of Tartu Functional Genomics Center, Zurich	454-GS-FLX- Titanium	2011
<i>Alcanivorax sp.</i> sk2-jrc		454-GS-FLX- Titanium	2011

The isolation and ecosystem characteristics of the sequenced genomes are presented in Table 4. All the four validly published strains were isolated from marine environment and have ecosystem annotation as “Environmental”. However, the isolation source of *Alcanivorax sp.* DG881 is “paralytic shellfish producing dinoflagellate *Gymnodinium catenatum* isolated from the Derwent estuary, Tasmania, Australia”.

Table 3. Isolation and ecosystem annotation of sequenced *Alcanivorax* genomes.

Genome	Isolation	Ecosystem
<i>Alcanivorax borkumensis</i> SK2	Seawater sediment sample in the Isle of Borkum, North Sea	Environmental
<i>Alcanivorax dieselolei</i> B5	oil-contaminated sea water at the Yellow River dock of Shengli oilfield	Environmental
<i>Alcanivorax hongdengensis</i> A-11-3	seawater from the Strait of Malacca	Environmental
<i>Alcanivorax pacificus</i> W11-5	Marine sediment, Pacific Ocean, China	Environmental
<i>Alcanivorax</i> sp.		
<i>Alcanivorax</i> sp. 43B_GOM-46m		
<i>Alcanivorax</i> sp. 97CO-5		
<i>Alcanivorax</i> sp. DG881	Paralytic shellfish producing dinoflagellate <i>Gymnodinium catenatum</i> isolated from the Derwent estuary, Tasmania, Australia	Host-associated
<i>Alcanivorax</i> sp. JRC		
<i>Alcanivorax</i> sp. sk2-jrc	Ispra, Italy	

The functional relatedness of the 10 *Alcanivorax* genomes based on the protein family annotations is presented in Figure 3. The topology of the functional relatedness is similar for the three species as in the phylogenetic relationship in Figure 1 where *A. borkumensis* and *A. hongdengensis* are recovered in the same clade/lineage whereas *A. dieselolei* was recovered in a distinct clade/lineage. Furthermore, the chromosomal map showing from outside to the center: Genes on forward strand (color by COG categories); Genes on reverse strand (color by COG categories); RNA genes (tRNAs green, rRNAs red, other RNAs black); GC content; and GC skew (Markowitz *et al.*, 2014) (Figure 4).

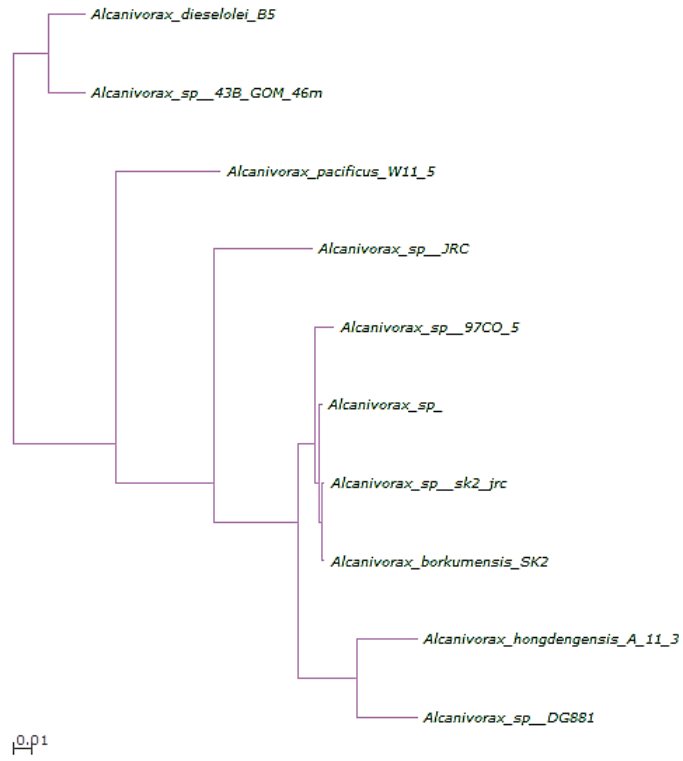
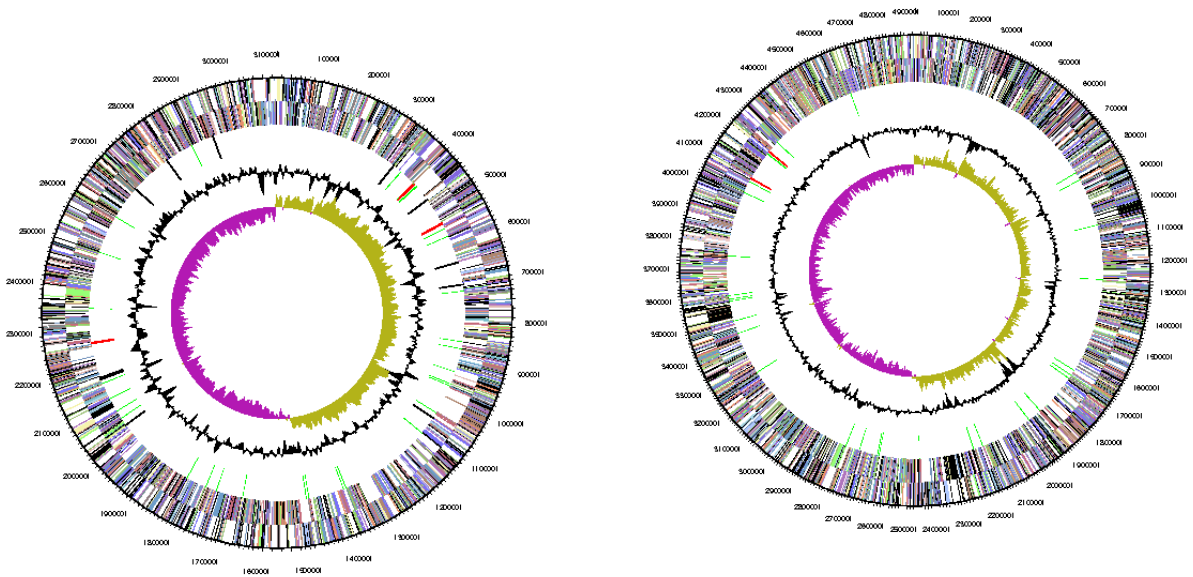


Figure 3. Functional relatedness of *Alcanivorax* genomes based on Pfam annotation of genes.



Alcanivorax borkumensis SK2

Alcanivorax dieselolei B5

Figure 4. Chromosomal maps of genomes of *Alcanivorax borkumensis* SK2 and *Alcanivorax dieselolei* B5.

Source: Chromosomal map was obtained from the integrated microbial Genomes System [<https://img.jgi.doe.gov/>].

2.3 *Alcanivorax* Genes for Habitat-Related Stress Response

The inventory of genes for habitat-related stress response systems in *Alcanivorax borkumensis* SK2 can be grouped into (i) genome stability, maintenance and DNA repair; (ii) chaperones; and (iii) detoxification of toxic compounds (Schneiker *et al.*, 2006). The subcategories of gene systems for this habitat-related response system are summarized in Table 4. Environmentally relevant phenotypes exhibited by *A. borkumensis* SK2 include biofilm formation, adaptation to UV exposure, and to growth at either low temperature or high salinity. Mini-Tn5 mutagenesis of the genome of *A. borkumensis* SK2 revealed mechanisms for the phenotypes including UV-induced modulations of c-di-GMP, transport of dicarboxylate, biosynthesis of extracellular polysaccharides, mRNA decay and modification, and a number of

novel stress-related regulatory systems (Sabirova *et al.*, 2008). A list of genes for UV response is presented in Table 5.

The role of the universal stress proteins has not been described for the *Alcanivorax* species. However, the USPs are known to be involved in UV tolerance and biofilm formation and maintenance (Boes *et al.*, 2006, Chen *et al.*, 2006, Mangalappalli-Illathu and Korber, 2006, Peters *et al.*, 2010, Schreiber *et al.*, 2006). For example, a USP of *Porphyromonas gingivalis* is involved in stress responses and biofilm formation (Figure 5) (Chen *et al.*, 2006).

Table 4. Habitat-related stress response systems in *Alcanivorax borkumensis* SK2

Genome stability, maintenance and DNA repair	Chaperones	Detoxification of toxic compounds
Photoreactivation	Protein secretion	Reduction and extrusion of arsenate
Mismatch repair	GroEL/GroES machinery	Mercury detoxification
Nucleotide excision repair	Prevention of the aggregation of newly synthesized proteins and unfolded proteins	Copper resistance
Base excision repair	ATP-dependent protease family	Export of heavy metals
SOS response	ATP-dependent Lon protease	
Recombinational repair	Ribosome-associated heat shock protein implicated in the recycling of the 50S subunit	
DNA alkylation damage repair	Cold shock proteins	
DNA ligases		

Source: (Schneiker *et al.*, 2006)

Table 5. The list of genes for DNA repair and cell division of *Alcanivorax borkumensis* SK2.

Gene name	Protein name
ABO_0427	UvrA, excision nuclease ABC, A subunit
ABO_0945	UvrB, excision nuclease ABC, B subunit
ABO_0945	UvrB, excision nuclease ABC, B subunit
ABO_1305	UvrC, excision nuclease ABC, C subunit
ABO_1801	RecA protein
ABO_0753	RuvB, Holliday junction DNA nuclease
ABO_2538	Rep ATP-dependent DNA helicase
ABO_1474	YfcB, site-specific DNA-methyltransferase
ABO_2735	ParA, parA family ATPase

Source: (Schneiker *et al.*, 2006). Genes found by transposon mutagenesis to be relevant for environmental adaptation to Ultra-Violet radiation.

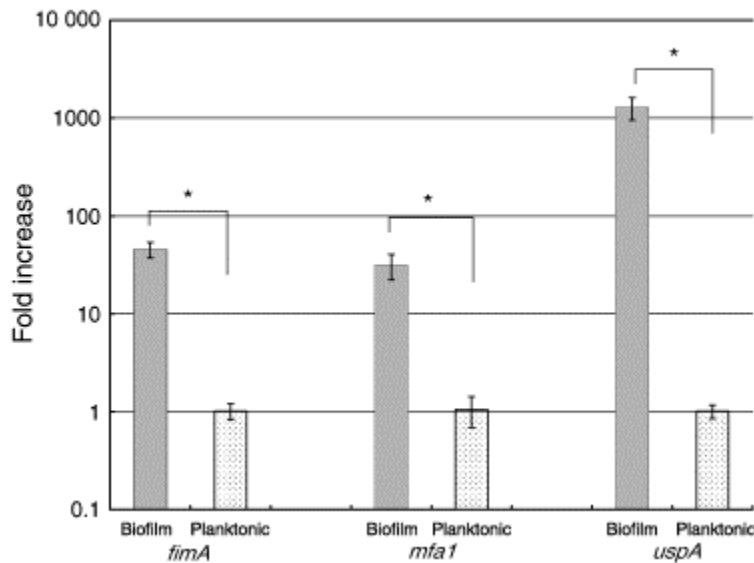


Figure 5. A universal stress protein of *Porphyromonas gingivalis* is involved in stress responses and biofilm formation

Biofilm formation is a form of cellular adaptation to environmental stress. There was increased expression of *uspA* during biofilm formation (1280-fold) which were significantly higher than other genes upregulated during biofilm formation. Source:(Chen *et al.*, 2006).

2.4 Universal Stress Protein Family

The universal stress protein (USP), originally named C13.5 on the basis of its migration across a two-dimensional IEF-PAGE gel, became of interest to investigators because of its underlying stimulation in response to a large variety of stress conditions. Various stress conditions include starvation for carbon, nitrogen, phosphate, sulfate, and amino acids and exposure to heat, oxidants and metals, uncoupler of the electron transport chain, polymyxin, cycloserine, ethanol and antibodies (Nachin *et al.*, 2005). Subsequently, the C13.5 protein has given its name to an orthologous group of proteins called the *UspA* superfamily of *Escherichia coli* in Frederick Neidhardt's laboratory and protein family (PF00582) in the Pfam database (Kvint *et al.*, 2003).

The conserved USP domain of 140-160 amino acids is prevalent in several domains of life including bacteria, archaea, fungi, flies and plants (Kvint *et al.*, 2003). However, it is not present in humans, which make it a prospective target for drug therapeutics (Hingley-Wilson *et al.*, 2010). Normally, organisms comprised of USP are equipped with duplicates of USP genes. In *E. coli*, there are six USP genes, *uspA*, *uspC*, *uspD*, *uspF*, *uspG* and *uspE* (tandem gene consisting of two USP domains) (Nachin *et al.*, 2005). Genes can encode either small Usp proteins (one domain), large Usp proteins (two tandem domains), or multiple domain proteins.

The corresponding Usp proteins can be divided into two sub-families based on sequence similarities. *UspA*, *UspC* and *UspD* belong to the *UspA* sub-family which does not bind ATP, whereas the *UspFG* sub-family (*UspF* and *UspD*) binds ATP. In tandem *UspE*, the first domain is more homologous to *UspA* sub-family and the second domain is closely related to *UspFG* sub-family (Heermann *et al.*, 2009). Based on structural analysis and their amino acid sequence, the Usp proteins have been divided into four different classes. In *E. coli*, *Usp A*, *UspC* (*yecG*) and

UspD (*yiiT*) belong to class I, *UspF* (*ynaF*) and *UspG* (*ybdQ*) belong to class II, and the two Usp domains of *UspE* (*ydaA*) separate into classes III and IV (Nachin *et al.*, 2005) (Figure 6).

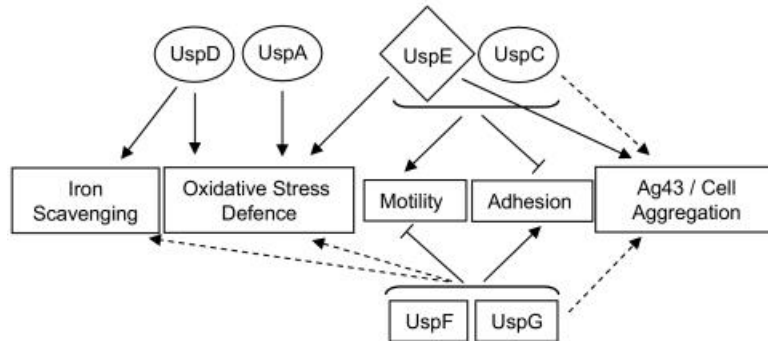


Figure 6. Role of the 6 *E. coli* Usps in oxidative stress defense, iron metabolism, and cell surface properties.

The name of each Usp is surrounded by a shape symbolizing the class it belongs to: a circle for class I, a square for class II, and a diamond for class III and IV. An arrow represents a positive effect of the Usp protein in a specific function, whereas a T shape signifies a negative effect. Major and minor effects of the Usps in the different functions are represented by solid and dashed lines, respectively. The brackets indicate that both of the included proteins are involved in the indicated process. For example, UspC and UspE both affect motility positively and adhesion negatively. Source: (Nachin *et al.*, 2005)

2.5 Role of *Alcanivorax* species in oil-spill clean-up

Contamination with petroleum and the consequences therein has been a common concern in oil-producing areas. Many halotolerant prokaryotes and eukaryotic degraders have been identified and reviewed by McGenity (McGenity, 2010). Among the hydrocarbon degrading bacteria, members of the genus *Alcanivorax* which are obligate hydrocarbon degraders and can utilize a wide range of crude oil compounds (Kasai *et al.*, 2002) tend to play a particular role in the clean-up of marine oil pollution (Cappello *et al.*, 2007).

In order to increase the natural populations of hydrocarbon-degrading bacteria, one of the techniques mainly used during the processes of bioremediation is the introduction of nutrients (e.g. nitrogen and phosphorus) for the development of the hydrocarbon-degrading bacteria. The presence of oil, together with nutrients availability, allows alive but inactive hydrocarbon-degrading bacteria to become active and, because of the presence of their sole source of carbon

(hydrocarbons), they become the dominant group of marine microbial community. The functional activation of specific catabolic pathways allows hydrocarbons degradation.

The growth of *Alcanivorax* in crude or heavy oil was not particularly rapid in comparison with other oil-degrading bacteria inhabiting seawater, although the role of *Alcanivorax* on bioremediation processes and the ability to become the most abundant species is still a matter of debate (Kasai *et al.*, 2002). Kasai and co-authors explained this phenomenon (i) by the ability of this bacterium to use a wide number of compounds of the oil as preferred energetic and nutritional source; and (ii) by the property of the biosurfactant of lipidic nature produced from *Alcanivorax* that increases the availability of hydrocarbons of the oil for the organism.

The biosurfactant of *Alcanivorax borkumensis* is one of the most efficient biosurfactants produced by bacteria; this facilitates emulsification, enhances bioavailability and therefore may have the potential to accelerate the degradation of other types of hydrophobic pollutants by bacteria or bacterial consortia. Thus, *Alcanivorax* not only degrades oil hydrocarbons in vitro, but seems to play a crucial role in the natural cleaning of oil-polluted marine systems. The biosurfactant of *Alcanivorax dieselolei* B-5, a species that uses diesel oil as the sole carbon and energy source, is a linear proline lipid biosurfactant monomer with a molecular mass of 355 (Figure 7). The proline lipid biosurfactant from strain B-5 has low toxicity, not membrane-bound, high efficiency with a low critical micelle concentration (CMC) value and stable in extreme temperature and pH (Qiao and Shao, 2010).

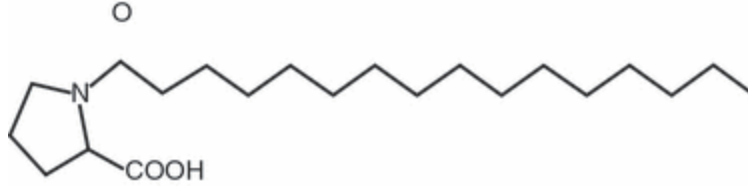


Figure 7. Deduced molecular structure of the lipopeptide produced by strain B-5.

The molecular structure of the monomer with a molecular mass of 355 was drawn using Chemdraw software version 8.0, based on PTA-AA (phenyl isothiocyanate-amino acid) and Mass Spectrometry results. The structure was proposed to be a proline and a C_{16:0} fatty acid, linked by an acylamide bond. Source: (Wu *et al.* 2009)

The unusual physiology and metabolic capability for hydrocarbon substrates, and the potential for biotechnological applications, make bacteria related to the *Alcanivorax* genus an interesting promise for bioremediation and lead to the basis of novel biotechnology strategies to accelerate the environmental remediation process (Cappello *et al.*, 2007). The morphology of *Alcanivorax borkumensis* at a water – n-hexadecane interface shows the presence of food (lipid) storage granules (Figure 8), a feature for coping with nutrient stress in marine environments (Santos *et al.* 2010; Kalscheuer *et al.* 2007).

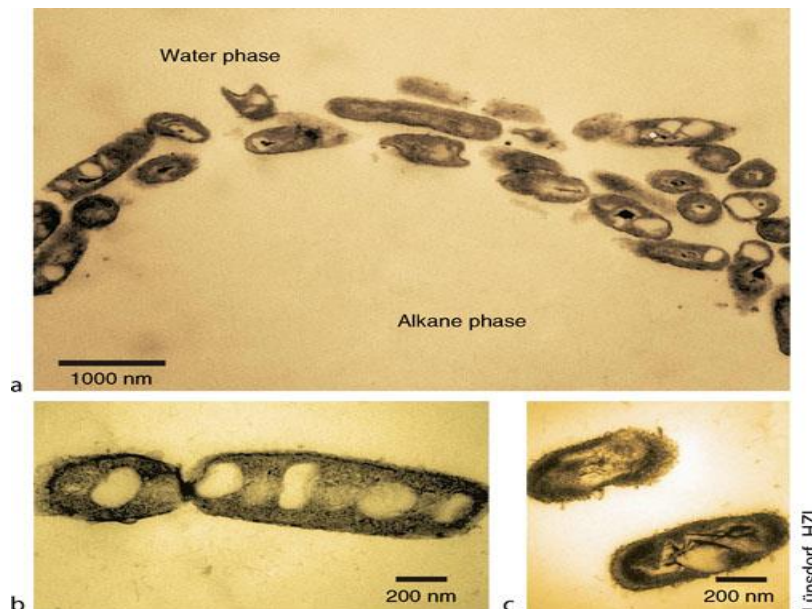


Figure 8. Transmission electron micrograph of *Alcanivorax borkumensis* cells growing at water – n-hexadecane interface.

The cellular shape is rather irregular and most cells contain electron-translucent inclusions, food storage granules, of different size and number (Credits: Heinrich Lümsdorf at the HZI). Source: (Martins dos Santos V. *et al.* 2010).

2.6 Bioinformatics Resources for Genomic Context Analytics

Integrated Microbial Genomes (IMG, <http://img.jgi.doe.gov/>) and BioCyc (<http://biocyc.org/>) are resources that aid in the visualization of chromosomal cassettes, gene context, and transcription units. The IMG data management system (Figure 9) provides analysis methods of gene context that are useful for the genomic structure and protein function prediction (Mavromatis *et al.*, 2009). It also provides the presence of fused genes, conservation of local neighbourhood of genes and co-occurrence of genes in genomes (Huynen *et al.*, 2000). BioCyc (Figure 10) makes a connection between the genes that are located in a local gene context and genes that are present in a transcription unit (Caspi *et al.*, 2014). A home page view of biocyc.org that includes view of the arrangement of genes in a comparative analysis of the genomic context.

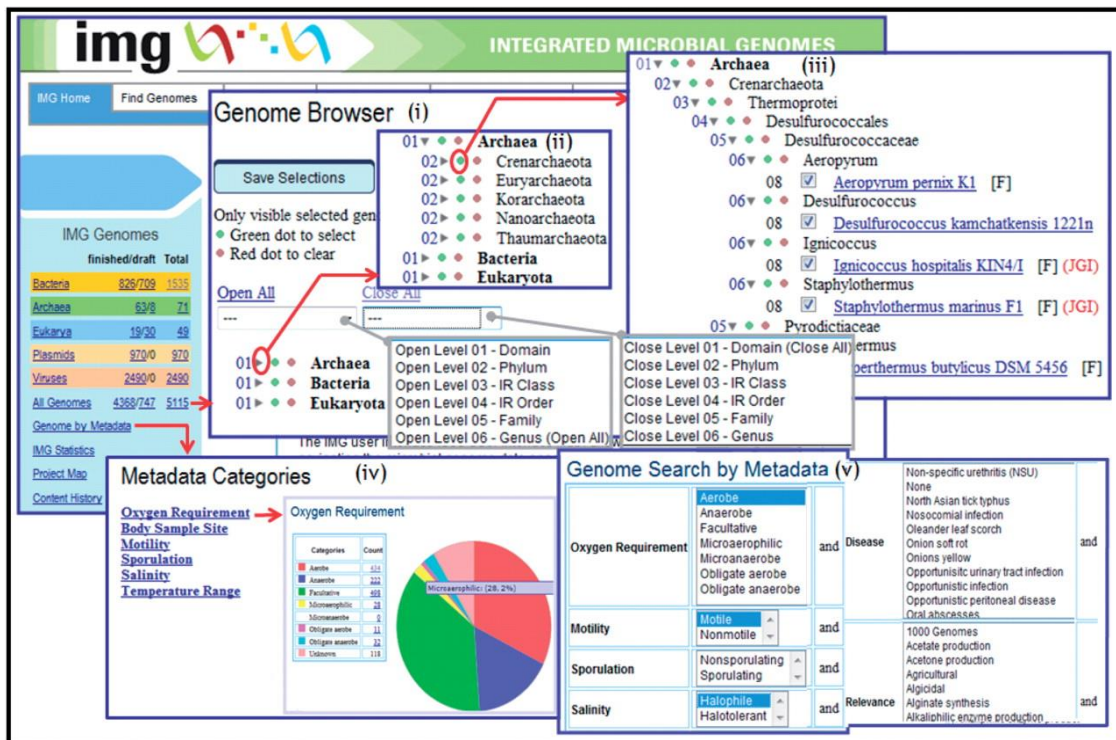


Figure 9. Overview of features of the Integrated Microbial Genomes (IMG) bioinformatics resource.

Source: (Mavromatis *et al.*, 2009)

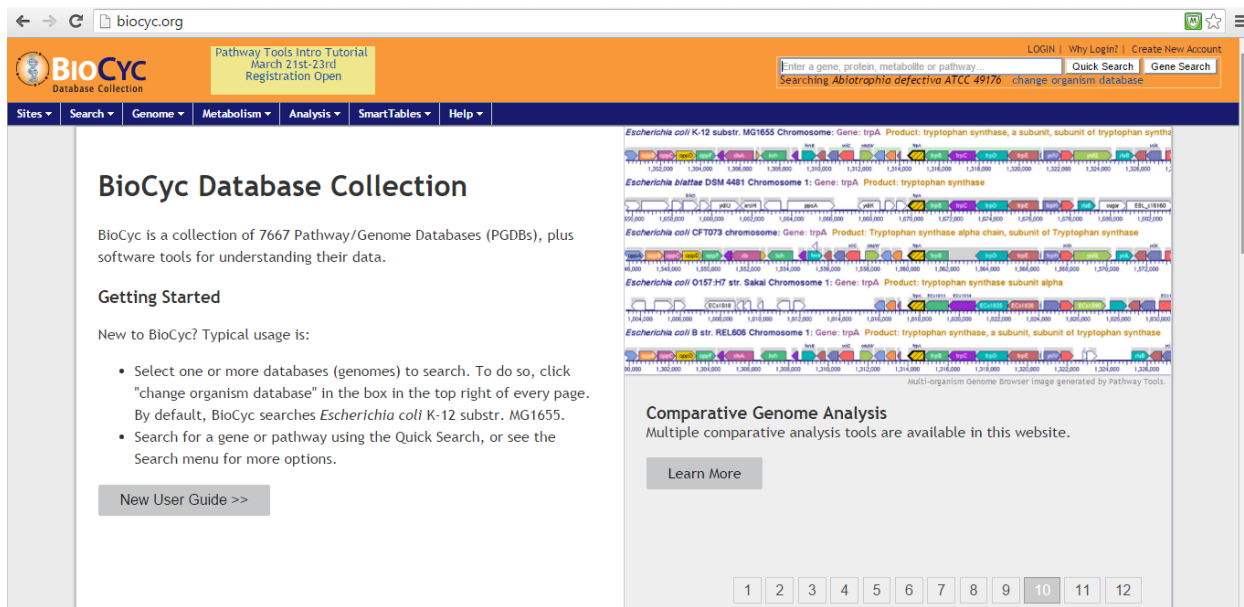



Figure 10. BioCyc database collection of pathway/genome databases and software tools.

Source: <http://biocyc.org/>

Another bioinformatics resource relevant to this project is STRING (Search Tool for the Retrieval of Interacting Genes/Proteins). STRING is a database of known and predicted protein interactions (Szklarczyk *et al.*, 2014) (Figure 11). The interactions include direct (physical) and indirect (functional) associations; they are derived from four sources: Genomic Context; High-throughput Experiments (Conserved) Co-expression and Previous Knowledge (Szklarczyk *et al.*, 2014). STRING has both network nodes and edges, the network nodes can either be coloured or white, while the edges represents predicted functional associations- the edges is usually drawn with 7 different lines which depicts different evidences (Figure 12) i.e. a green line shows neighbourhood evidence; a blue line shows co-occurrence; a red line shows the presence of fusion evidence; purple line shows experimental evidence, yellow line shows text mining evidence; black line shows co-expression evidence and light blue line shows database evidence (Szklarczyk *et al.*, 2011).

Home · Download · Help · My Data 

STRING - Known and Predicted Protein-Protein Interactions

protein name: (examples: #1 #2 #3)

(STRING understands a variety of protein names and accessions; you can also try a [random entry](#).)





organism:

interactors wanted:
 COGs Proteins

please enter your protein of interest...

What it does ...

STRING is a database of known and predicted protein interactions. The interactions include direct (physical) and indirect (functional) associations; they are derived from four sources:

Genomic Context 	High-throughput Experiments 	(Conserved) Coexpression 	Previous Knowledge 
--	---	---	---

STRING quantitatively integrates interaction data from these sources for a large number of organisms, and transfers information between these organisms where applicable. The database currently covers 9'643'763 proteins from 2'031 organisms.

STRING (*Search Tool for the Retrieval of Interacting Genes/Proteins*) is being developed at [CPR](#), [EMBL](#), [SIB](#), [KU](#), [TUD](#) and [UZH](#).
 STRING references: [Szklarczyk et al. 2015](#) / [2013](#) / [2011](#) / [2009](#) / [2007](#) / [2005](#) / [2003](#) / [Snel et al. 2000](#).
 Miscellaneous: [Access Statistics](#), [Robot Access Guide](#), [Supported Browsers](#).

What's New? This is version 10 of STRING - now covering more than 2000 organisms, and with improved prediction algorithms!
Sister Projects: check out [STITCH](#) and [eggNOG](#) - two sister projects built on STRING data!
Previous Releases: Trying to reproduce an earlier finding? Confused? Refer to our [old releases](#).

Figure 11. STRING (Search Tool for the Retrieval of Interacting Genes/Proteins)

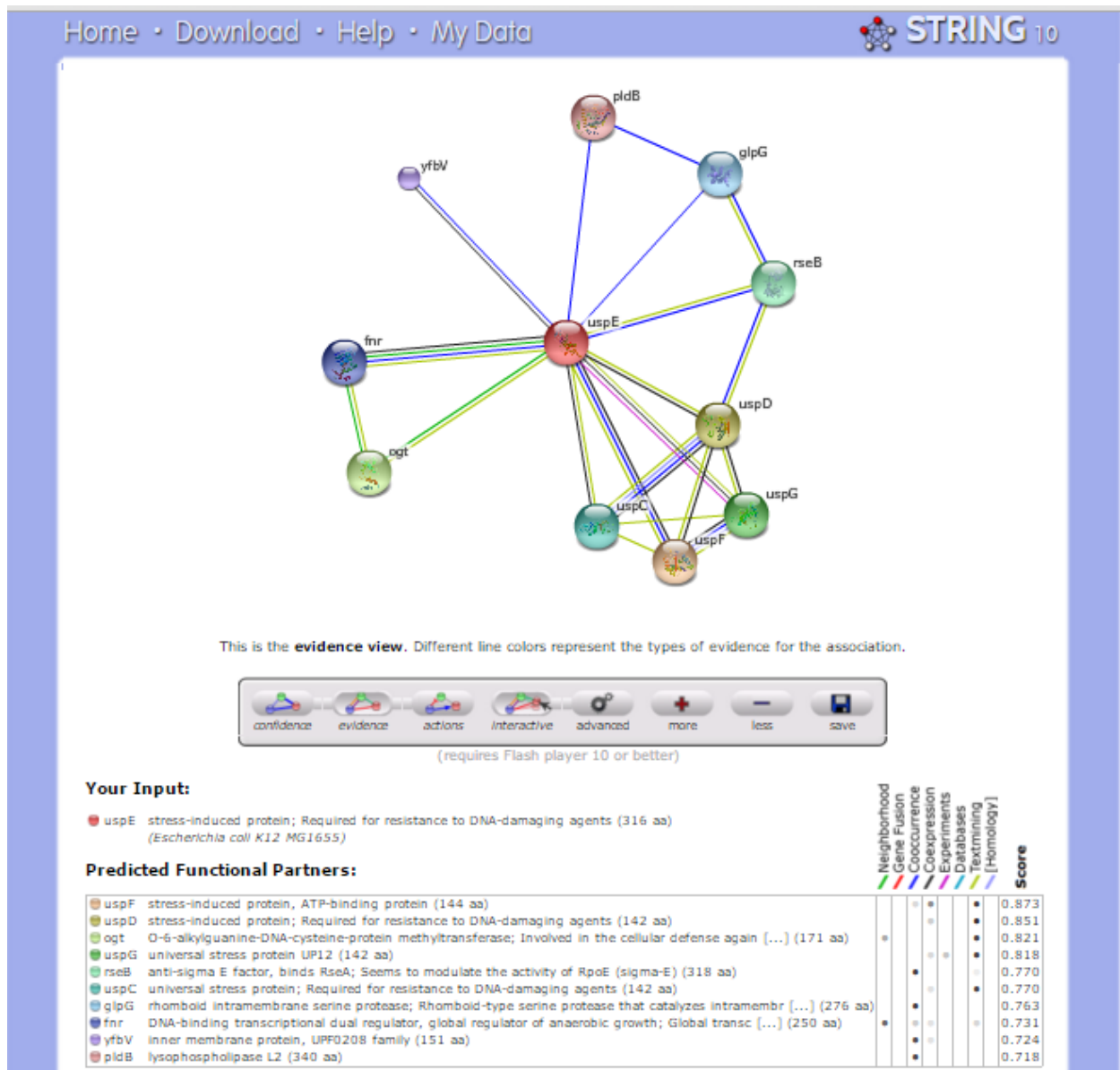


Figure 12. Results page of search for functional interactions of a protein in STRING database.

2.7 Visual Analytics and Bioinformatics

It is increasingly recognized that the study of biological systems is associated with volumes of data in which the acquisition of expertise in database, workflow, visualization and cloud computing techniques are necessary to make scientific progress in biology (Cottingham, 2012). The integrated approach of visual analytics combines fields such as visualization, human factors and data analysis, which in turn integrate different methodologies. Visual analytics

involves exploring, cleaning, gaining confidence in, summarizing, pursuing inconclusive paths, confirming facts and presenting findings about the data (Chabot, 2009, Chang *et al.*, 2009, Keim *et al.*, 2008). Visual analytics environments include software, hardware, computing power and display technology that enables the focus areas of visual analytics to be accomplished (Scholtz, 2006). The emphasis of this research is on the use of easy to deploy visual analytics software that includes ability to share visualization via web interfaces as well as conduct visual analytics tasks with a desktop version.

Visual analytics is a new emerging field of research and is recognized as the science of analytical reasoning by interactive visual interfaces (Thomas and Cook 2006). The interdisciplinary field is composed of research areas such as visualization, data mining, data management, data fusion, statistics and cognitive science (Keim, 2012).

The increasing amount of data availability and need for faster processing capabilities has given recognition to the software tools for visual analytics. Data is imported into the software and easily manipulated to produce representations and patterns. The constructed models are developed through the combination of electronic data processing and human ability to draw inferences.

Visual analytics is significant in the processing and analysis of large information enabling humans to gain an understanding, reasoning, and decision making. Typically, visual analytics tools enable the import of multiple files with similar data content to be linked and manoeuvred to give various views. There are a number of available tabs to access a new screen without deleting the information created. The data is divided into a dimension and measure section but is interchangeable. These fields have filters, allowing the selection of interest to narrow the focus

of research. Bar/Line graphs, pie charts, and tables are designed to show the relationship of the data that is researched. These illustrations consist of the colours for differentiation, if needed. Calculations are also obtainable through the visual analytics tool. Once the desired image and with efficient results have been achieved, it can be exported for substantial analysis. Tableau Software (<http://www.tableau.com>) is an example of visual analytics software that provides features to help users understand their data. Visualizations can also be shared through the Tableau Public website. The interface for interacting with features of Tableau is shown in Figure 13.

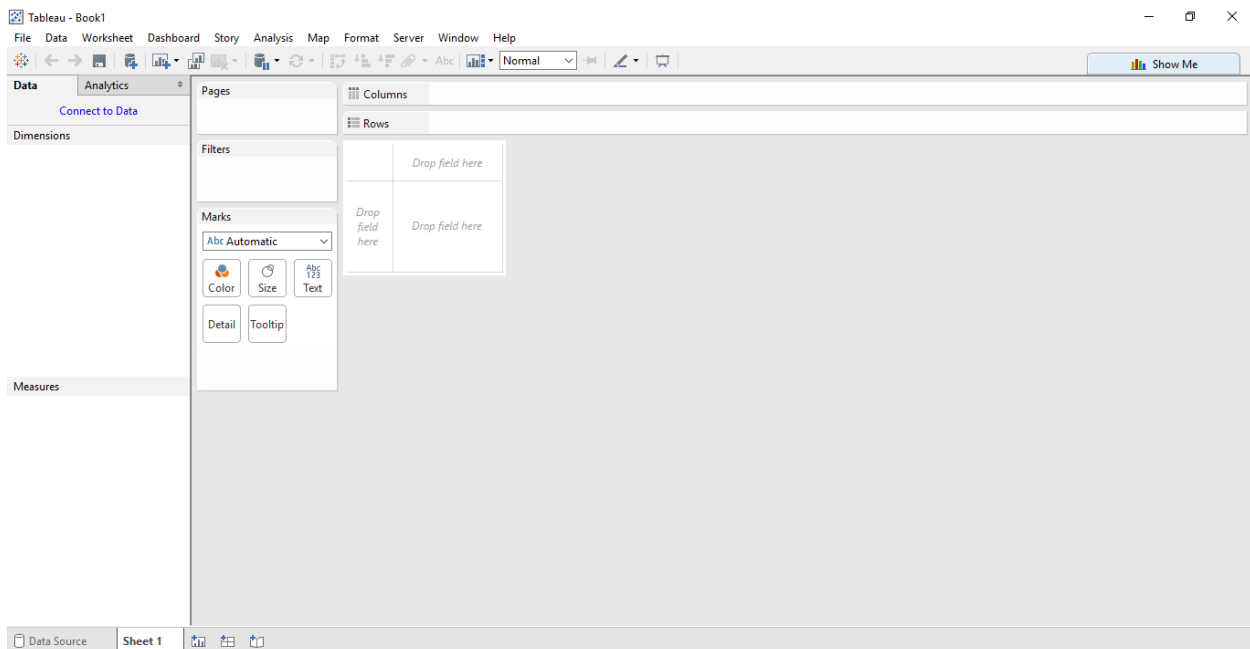


Figure 13. Tableau Software user interface for conducting visual analytics.

Visual analytics is divided into different categories depending on how the information is represented. The presentation of the data can be categorized as characteristics with values and structure with relationship, which can affect the decision-making process. The cognitive behaviour of the user is also important when constructing and observing the views to pose an analysis.

Therefore, through visual analytics software and hardware infrastructure, it is possible to integrate results from bioinformatics software as well as provides models of how the data are related. In this present study, visual analytics was used to integrate functional annotation data obtain from multiple data sources to gain new knowledge on the functions of universal stress proteins. Visual analytics procedures using Tableau Software have been applied to data on universal stress proteins from species of the genus *Bacillus* and *Schistosoma* (Isokpehi *et al.* 2011; Williams *et al.* 2012; Mbah *et al.* 2013). A web-based resource for Schistosoma USP is shown in Figure 14.

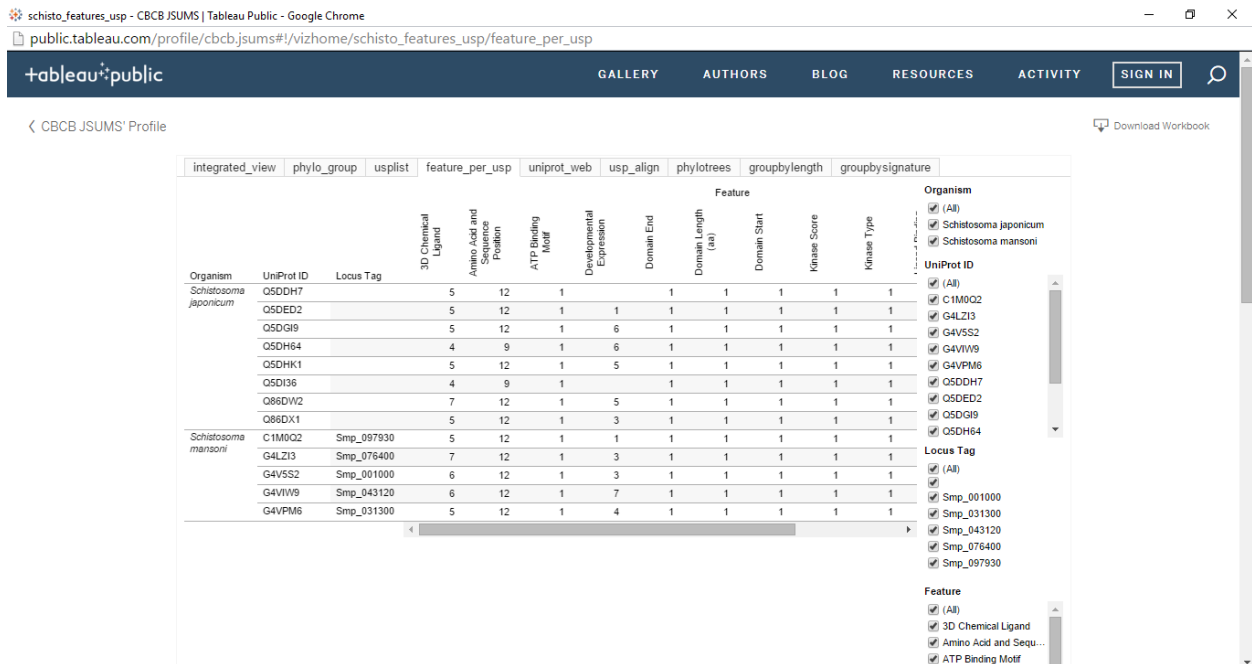


Figure 14. Web-based resource for integrating annotation features for universal stress proteins of *Schistosoma mansoni* and *Schistosoma japonicum*.

Link: [http://public.tableau.com/profile/cbcb.jsums#!/vizhome/schisto_features_esp/feature_per_esp](http://public.tableau.com/profile/cbcb.jsums#!/vizhome/schisto_features_usp/feature_per_esp)

CHAPTER 3

MATERIALS AND METHODS

The materials and methods followed in this thesis involved several methods for the functional annotation analytics of proteins developed by Isokpehi and colleagues (Isokpehi, *et al.*, 2011; Williams *et al.*, 2012). The integrated bioinformatics and visual analytics methods allows researchers to gain insights and contribute to the knowledge on the function of universal stress proteins in diverse organisms. The overview of the research approach for elucidating the functions of *Alcanivorax* universal stress proteins is presented in Figure 15.

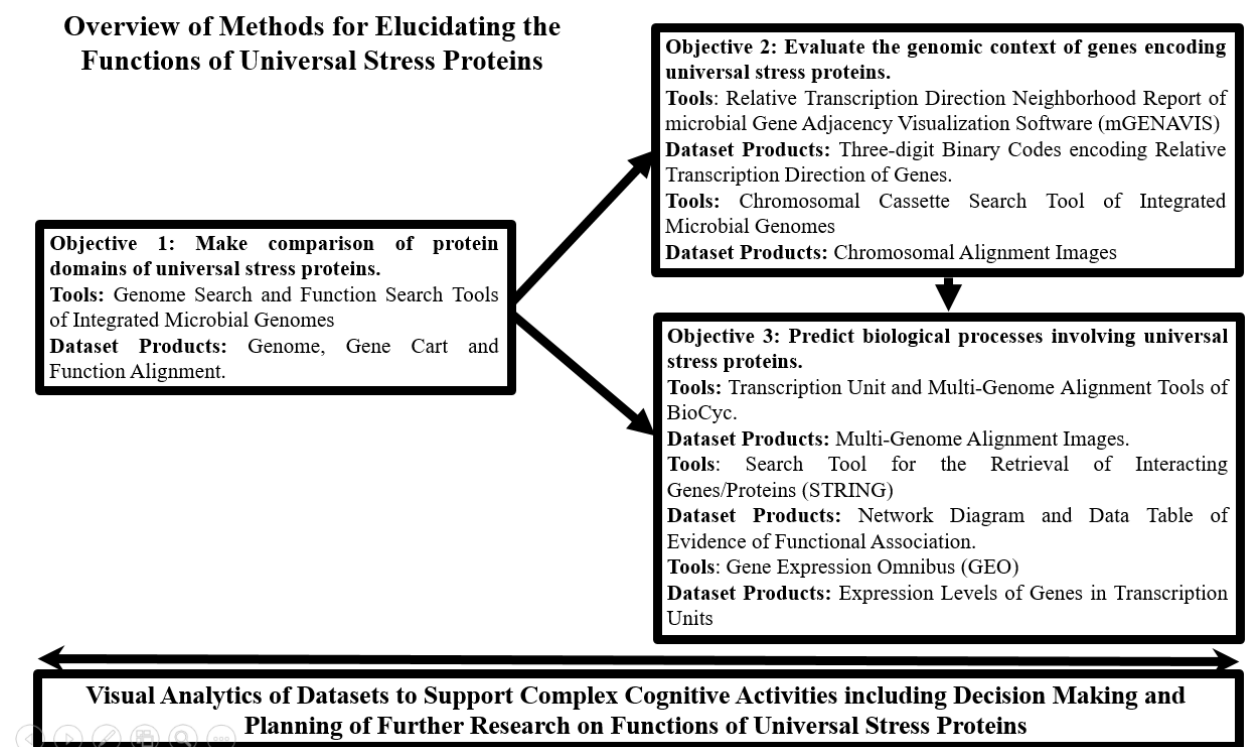


Figure 15. Overview of research approach for elucidating functions of universal stress proteins.

3.1 Objective 1: To make comparison among the protein domain organization of universal stress proteins encoded in *Alcanivorax* genomes.

3.1.1 Introduction

The protein domain organization provided information on the functions of a protein sequence. This objective was designed to reveal the number of USP domains per protein as well as other protein domains that are present. Furthermore, evaluation of the coordinates of the protein domains will identify shared and divergent protein domain coordinates.

3.1.2 Retrieval of Datasets

Data sets for protein domain organization were obtained from the Integrated Microbial Genomes (IMG) system (Mavromatis *et al.*, 2009). The Genomes Search Tool was used to obtain the annotation data for the genomes of the *Alcanivorax* in the IMG system. The data field categories are Genome Field, Project Metadata and Data Statistics. The Genome Field category included columns such as Domain, Status, Study Name and Genome Name/Sample. This data set is designated as Genome Data set.

The Function Search Tool in the IMG system was used to retrieve genes encoded with the Pfam domain pfam00582 (universal stress protein family) in the *Alcanivorax* genomes. The genes were added to the Gene Cart where the following data columns: Gene ID, Locus Tag, Gene Product Name, Genome and other data field categories (Gene Field, Scaffold/Contig Field and Function Category) were exported to a spread sheet file. The Function Category included a column labelled Pfam, which provided data on the protein domain present in the USP. This data set was designated Gene Cart data set.

The IMG Gene Cart contains the following tabs: Genes in Cart; Functions, Upload & Export; Chromosome Map, Sequence Alignment; Gene Neighbourhoods and Profile & Alignment. Function Alignment button in the Profile & Alignment tab leads to the data sets for the COG (Cluster of Orthologous Groups of Proteins); KOG (Eukaryotic Orthologous Groups) and Pfam (Protein family). The Pfam function alignment on the protein was selected and the data columns exported to a spread sheet file. The data fields were Gene ID, Pfam ID, Pfam Name, Percent Alignment on Query Gene, Query Start, Query End, Alignment On Query Gene, HMM (Hidden Markov Model) Score and Genome. This data set was designated Function Alignment data set.

3.1.3 Visual Analytics of Protein Domain Data

Visual analytics tasks were conducted using the Tableau Desktop Professional (Tableau Software WA, USA). The Gene Cart and Functional Alignment data sets were uploaded to Tableau Desktop Professional (Tableau Software, WA, USA) for visual analytics tasks. There were shared fields that allowed integration of data sets. Several visual analytics view designs were developed to gain deeper insights into the protein domain organization of the USPs encoded in the *Alcanivorax* genomes.

3.2 Objective 2: To evaluate the genomic context of genes encoding the universal stress protein

3.2.1 Introduction

The genomic context of a gene of interest can be defined as the chromosomal cassette that contains the gene. In the IMG system, a chromosomal cassette is defined as a stretch of protein coding genes with intergenic distance smaller or equal to 300 base pairs. In addition, a chromosomal alignment search in the IMG system retrieves genomic regions in other genomes

that are similar to the search genomic region. The result of the search provided only statistically significant alignments in the IMG database. Results also included images as well as web page text (html codes) that described each chromosomal alignment. These data sets provided opportunities for evaluation of the genomic context by a researcher to determine noteworthy and potentially biological significant findings. In some instances it is impractical to review the genomic context of all genes retrieved therefore representative sequences are selected on basis of characteristics such as relative transcription direction of gene neighbourhood, common protein length and sequence clustering after multiple sequence alignment. The relative transcription direction was determined using the microbial Gene Adjacency Visualization Software (mGENAVIS <http://bioinformatics.ui.edu.ng/geneadjacency/>).

3.2.2 Relative Transcription Direction of Genes for *Alcanivorax* Universal Stress Proteins

The software mGENAVIS was used to compare all the *Alcanivorax* genomes in an archive of the genome annotation files (ftp://ftp.patricbrc.org/patric2/patric2_archive/genomes/). The mGENAVIS encodes and visualizes the relative transcription direction of genes in a microbial genome. The data file generated from the genome annotation file (RefSeq.cds.tab) contains the relative transcription direction (forward [+] or reverse [-] DNA strand) of each gene including those for the universal stress proteins. The types of relative transcription direction encoding are 111, 110, 011 and 010. Codes 111, 110 and 011 could be part of operons.

3.2.2 Chromosomal Cassette Search

The chromosomal cassette search link for a gene was selected from the gene page in Integrated Microbial Genomes (IMG) resource. A link is https://img.jgi.doe.gov/cgi-bin/edu/main.cgi?section=GeneCassette&page=geneCassette&gene_oid=2546926451&type=pfa. The link will display chromosomal alignment for gene 2546926451 in the IMG resource. The

resulting page was inspected for noteworthy and potentially biological significant findings from the cassettes and alignments. Observations of interest were: (1) the organism source of genomes reported in top three genomic regions; (2) the presence in the alignment of genome from other bacteria or archaea taxa; (3) the taxonomic lineage of all the genomes retrieved by the chromosomal cassette search. These observations could indicate lineage association of the genomic context of the gene of interest.

3.3 Objective 3: To predict biological processes involving *Alcanivorax* universal stress proteins

3.3.1 Evaluation of Stress Response Equipped Transcription Units

The Locus Tag of the gene of interest was searched in the genome source database in BioCyc [<http://biocyc.org/>]. The status of the gene as a multi-gene transcription unit was noted (0 for absence; 1 for present). This encoding was used to group the response to stress genes being compared. Thus, this encoding was determined by grouping the USP genes of *Alcanivorax* into two groups.

For each representative USP gene, the genomic context alignment with orthologs in selected genomes was determined with the BioCyc Multi-Genome Alignment Tool. Noteworthy findings from visual inspection of the alignment were reported. For each representative gene, the Locus Tag was used to search the STRING database (Szklarczyk *et al.*, 2011). The STRING network with the types of evidence of functional association was stored as part of the knowledge set to make decision on molecular and phenotypic characterization of stress response equipped biological processes. The confidence of the interactions was assessed using the combined score (low confidence: scores <0.4; medium: 0.4 to 0.7; high: >0.7).

3.3.2 Gene Expression Patterns of Transcription Units Encoding *Alcanivorax* Universal Stress Proteins

The Gene Expression Omnibus (GEO) resource was searched for gene expression (microarray) datasets that could be used to determine the expression patterns of genes for universal stress protein and their partner genes in transcription units. The expression levels of the genes of interest were retrieved and box plots were constructed with Tableau to determine the patterns of the gene expression. Genes in transcription unit are expected to have similar patterns of gene expression levels.

CHAPTER 4

RESULTS

This chapter presents results obtained from bioinformatics analysis of *Alcanivorax* genomes and visual analytics procedures implemented on the results of bioinformatics analysis. The headings of the findings are presented in Table 6.

Table 6. Findings from research project

Objective	Results
1	Genome Statistics of <i>Alcanivorax</i> in the Integrated Microbial Genomes Database
1	Gene count and protein domain organization of universal stress proteins encoded in <i>Alcanivorax</i> genomes
1	Protein sequence length and alignment of universal stress protein domain
2	Relative Transcription Direction of <i>Alcanivorax</i> Universal Stress Protein Genes
2	Chromosomal Cassette Search
3	Evaluation of Stress Response Equipped Transcription Units
3	Prediction functional partners to <i>Alcanivorax borkumensis</i> universal stress protein

4.1 Objective 1: To make comparison among the protein domain organization of universal stress proteins encoded in *Alcanivorax* genomes.

4.1.1 Genome Statistics of *Alcanivorax* in the Integrated Microbial Genomes Database

A total of 15 genome information datasets of *Alcanivorax* were retrieved from the Integrated Microbial Genomes database version 4.530 June 2015 (<http://img.jgi.doe.gov/edu/>). In Table 7 (order by GC%), the following data are presented for the *Alcanivorax* genomes: the Genome Name, IMG Genome Identifier, Genome Size, Gene Count, and Guanosine + Cytosine (G+C) content (also GC%). Of the genomes, genome of *Alcanivorax sp.* sk2-jrc with genome identifier 2510461004 had the highest gene count of 6479. *Alcanivorax borkumensis* SK2 (Genome ID: 637000004) had the lowest gene count of 2817. *Alcanivorax borkumensis* SK2 with IMG Genome ID 2623620884 with gene count of 2880 replaced the older version that had IMG identifier: 637000004. The GC% ranged from 55% to 63% with an average of 58% for the

sequenced *Alcanivorax* genomes. The following genomes had GC% greater than upper quartile of GC% of 60%: *Alcanivorax dieselolei* B5 (62%), *Alcanivorax hongdensis* A-11-3 (61%), *Alcanivorax pacificus* W11-5 (63%) and *Alcanivorax* sp. 43B GOM-46m (61%).

Table 7. Genomes of *Alcanivorax* in the Integrated Microbial Genomes database.

Genome Name	IMG Genome ID	Genome Size	Gene Count	GC (%)
<i>Alcanivorax borkumensis</i> SK2	637000004	3120143	2817	55
<i>Alcanivorax borkumensis</i> SK2	2623620884	3120143	2880	55
<i>Alcanivorax</i> sp. 97CO-5	2568526460	3251558	3027	55
<i>Alcanivorax</i> sp. JRC	2504643016	3116196	5103	55
<i>Alcanivorax</i> sp. Rast	2507149004	3111156	2973	55
<i>Alcanivorax</i> sp. sk2-jrc	2507149008	3111156	2895	55
<i>Alcanivorax</i> sp. sk2-jrc	2510461004	3102433	6479	55
<i>Alcanivorax</i> sp. 19-m-6	2600255028	4132804	3874	56
<i>Alcanivorax jadensis</i> T9	2600255029	3629371	3337	58
<i>Alcanivorax</i> sp. DG881	647533109	3804728	3384	58
<i>Alcanivorax</i> sp. DSM 26293	2634166299	3706254	3475	59
<i>Alcanivorax hongdengensis</i> A-11-3	2531839231	3664876	3459	61
<i>Alcanivorax</i> sp. 43B_GOM-46m	2546825531	4743861	4422	61
<i>Alcanivorax dieselolei</i> B5	2521172691	4928223	4470	62
<i>Alcanivorax pacificus</i> W11-5	2537562047	4137438	3804	63

4.1.2 Gene count and protein domain organization of universal stress proteins encoded in

Alcanivorax genomes.

A total of 62 genes encoding the universal stress protein domain (pfam00582) were associated with the 15 *Alcanivorax* genomes (Table 8). Eleven of the 15 genomes have a USP gene count of 4. Both genomes of the *A. borkumensis* SK2 had 4 USP genes. The following genomes had five USP genes: *Alcanivorax jadensis* T9, *Alcanivorax* sp. 43B_GOM-46m, and *Alcanivorax* sp. sk2-jrc (Genome ID: 2510461004).

Table 8. Count of genes in *Alcanivorax* genomes in the Integrated Microbial Genomes database.

Genome Name	IMG Genome ID	Count of USP Genes*
<i>Alcanivorax sp.</i> DG881	647533109	3
<i>Alcanivorax borkumensis</i> SK2	637000004	4
<i>Alcanivorax borkumensis</i> SK2	2623620884	4
<i>Alcanivorax dieselolei</i> B5	2521172691	4
<i>Alcanivorax hongdengensis</i> A-11-3	2531839231	4
<i>Alcanivorax pacificus</i> W11-5	2537562047	4
<i>Alcanivorax sp.</i> 19-m-6	2600255028	4
<i>Alcanivorax sp.</i> 97CO-5	2568526460	4
<i>Alcanivorax sp.</i> DSM 26293	2634166299	4
<i>Alcanivorax sp.</i> JRC	2504643016	4
<i>Alcanivorax sp.</i> Rast	2507149004	4
<i>Alcanivorax sp.</i> sk2-jrc	2507149008	4
<i>Alcanivorax jadensis</i> T9	2600255029	5
<i>Alcanivorax sp.</i> 43B_GOM-46m	2546825531	5
<i>Alcanivorax sp.</i> sk2-jrc	2510461004	5

* Genes with Universal Stress Protein (USP) Domain

An overview (Figure 16) of the protein domain organization annotation for the 62 USP genes showed that in addition to the USP domain (pfam00582), gene 2546926451 (Locus Tag: N537DRAFT_02674) of *Alcanivorax sp.43_GOM-46m* had annotations for pfam00512 [His Kinase A (phospho-acceptor) domain], pfam02518 [Histidine kinase, DNA gyrase B, and HSP90-like ATPase], pfam02702 (Osmosensitive K⁺ channel His kinase sensor domain) and pfam13493 [Domain of unknown function (DUF4118)]. The *Alcanivorax* genomes retrieved have a collection of genes encoding proteins that have single USP domain and those that have two USP domains (Figure 16). The four *Alcanivorax borkumensis* SK2 genomes encode two each of single (ABO_1011, ABO_2141) and double domains (ABO_1340, ABO_1511) universal stress proteins.

4.1.3 Protein sequence length and alignment of universal stress protein domain

The distribution of the amino acid length of the 62 *Alcanivorax* USPs revealed several patterns of common and distinguishing features for genomes. The 23 amino acids sequence lengths types based on number of amino acids for *Alcanivorax* USPs were 33, 37, 61, 104, 139, 145, 146, 148, 150, 152, 157, 230, 237, 278, 279, 280, 282, 284, 286, 290, 305, 310, and 880 (Figure 17). Sequence lengths types of 146, 279 and 305 amino acids had frequency of 9, 7 and 11 respectively. The sequence length annotations were identical for the two *Alcanivorax borkumensis* SK2 genomes. However, the two genomes tagged with strain sk2-jrc (Genome IDs: 2507149008 and 2510461004) did not have identical sequence length annotations. In particular, Genome 2510461004 had predicted sequence lengths of 33, 37, 61 and 104 amino acids.

The 62 USP sequences were grouped by Pfam ID and the Percent Alignment on Query Gene (Figure 18). This grouping view also included the query start and query end values of the domain organisation. Protein sequences from different genomes with identical query start and query end values are predicted possible orthologs. For example, a group of 6 genes (638079441 [ABO_1340], 2504699500 [Alc_pool_00025010], 2507173281, 2507185410 [EKC_04173], 2570067287 [Y017_00170] and 2626078400 [Ga0076350_111413]) encoding protein sequence length of 350 aa had a 45.9% alignment on the protein sequence. The second USP domain of the 305aa type USP overlaps 48.52% from position 148aa to position 295aa. *Alcanivorax dieselolei* and *Alcanivorax sp. 43B_GOM-46m* had identical positions for query start and query end coordinates for the USP domain in three instances of the percent alignments 45.1%, 45.57% and 45.8% (Figure 18).

Genome	Gene ID	Locus Tag	Pfam ID					
			pfam00512	pfam00582	pfam02518	pfam02702	pfam13483	
Alcanivorax borkumensis SK2	638079104	ABO_1011			1			
	638079441	ABO_1340			2			
	638079615	ABO_1511			2			
	638080255	ABO_2141			1			
	2626078056	Ga0076350_111069			1			
	2626078400	Ga0076350_111413			2			
	2626078582	Ga0076350_111595			2			
	2626079239	Ga0076350_112253			1			
Alcanivorax dieselolei B5	2521960177	B5T_01588			2			
	2521960410	B5T_01819			1			
	2521960869	B5T_02274			2			
	2521961056	B5T_02456			2			
Alcanivorax hongdengensis A-11-3	2532721685	A11A3_04780			1			
	2532722183	A11A3_07248			1			
	2532722799	A11A3_10316			2			
	2532723912	A11A3_15846			2			
Alcanivorax jadensis T9	2600659746	Ga0060136_00014			2			
	2600659996	Ga0060136_00264			2			
	2600660338	Ga0060136_00606			1			
	2600660675	Ga0060136_00944			1			
	2600661501	Ga0060136_01770			1			
Alcanivorax pacificus W11-5	2539615017	S7S_00941			2			
	2539615329	S7S_01246			1			
	2539616560	S7S_02463			2			
	2539617055	S7S_02956			2			
Alcanivorax sp. 19-m-6	2600655887	Ga0060137_00029			2			
	2600656145	Ga0060137_00287			1			
	2600656948	Ga0060137_01090			2			
	2600658412	Ga0060137_02554			1			
Alcanivorax sp. 43B_GOM-46m	2546925717	N537DRAFT_01940			1			
	2546925967	N537DRAFT_02190			2			
	2546926451	N537DRAFT_02674	1	1	1	1	1	
	2546926495	N537DRAFT_02718			2			
	2546926673	N537DRAFT_02896			2			
Alcanivorax sp. 97CO-5	2570067287	Y017_00170			2			
	2570067634	Y017_01905			1			
	2570068302	Y017_10235			2			
	2570069632	Y017_04425			1			
Alcanivorax sp. DG881	647581859	ADG881_755			1			
	647582196	ADG881_1440			2			
	647582426	ADG881_443			2			
Alcanivorax sp. DSM 26293	2635087216	Ga0074803_101101			1			
	2635088066	Ga0074803_102238			2			
	2635088286	Ga0074803_102458			2			
	2635089429	Ga0074803_10626			1			
Alcanivorax sp. JRC	2504697814	Alc_pool_00008150			1			
	2504699500	Alc_pool_00025010			2			
	2504699825	Alc_pool_00028260			1			
	2504699826	Alc_pool_00028270			1			
Alcanivorax sp. rast	2507172017	Null			1			
	2507173100	Null			2			
	2507173281	Null			2			
	2507173515	Null			1			
Alcanivorax sp. sk2-jrc	2507184742	EKC_00850			1			
	2507185227	EKC_03273			2			
	2507185410	EKC_04173			2			
	2507185565	EKC_04944			1			
	2510478651	Alc18_00019270			1			
	2510478652	Alc18_00019280			1			
	2510479451	Alc18_00027270			1			
	2510479864	Alc18_00031400			1			
	2510479869	Alc18_00031450			1			

Figure 16. Pfam content of universal stress proteins in *Alcanivorax* genomes.

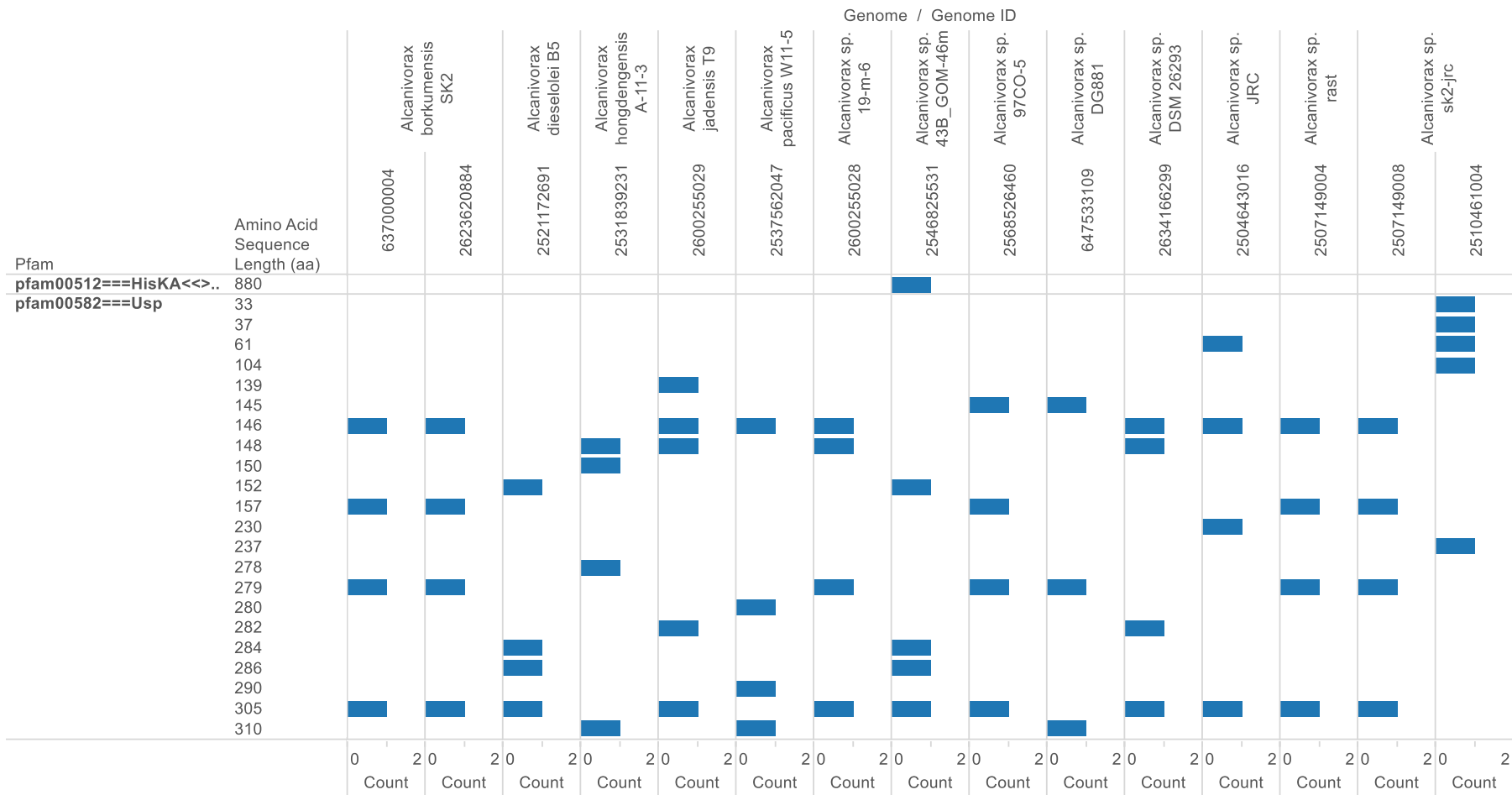


Figure 17. Distribution of amino acid sequence lengths of the *Alcanivorax* universal stress proteins.

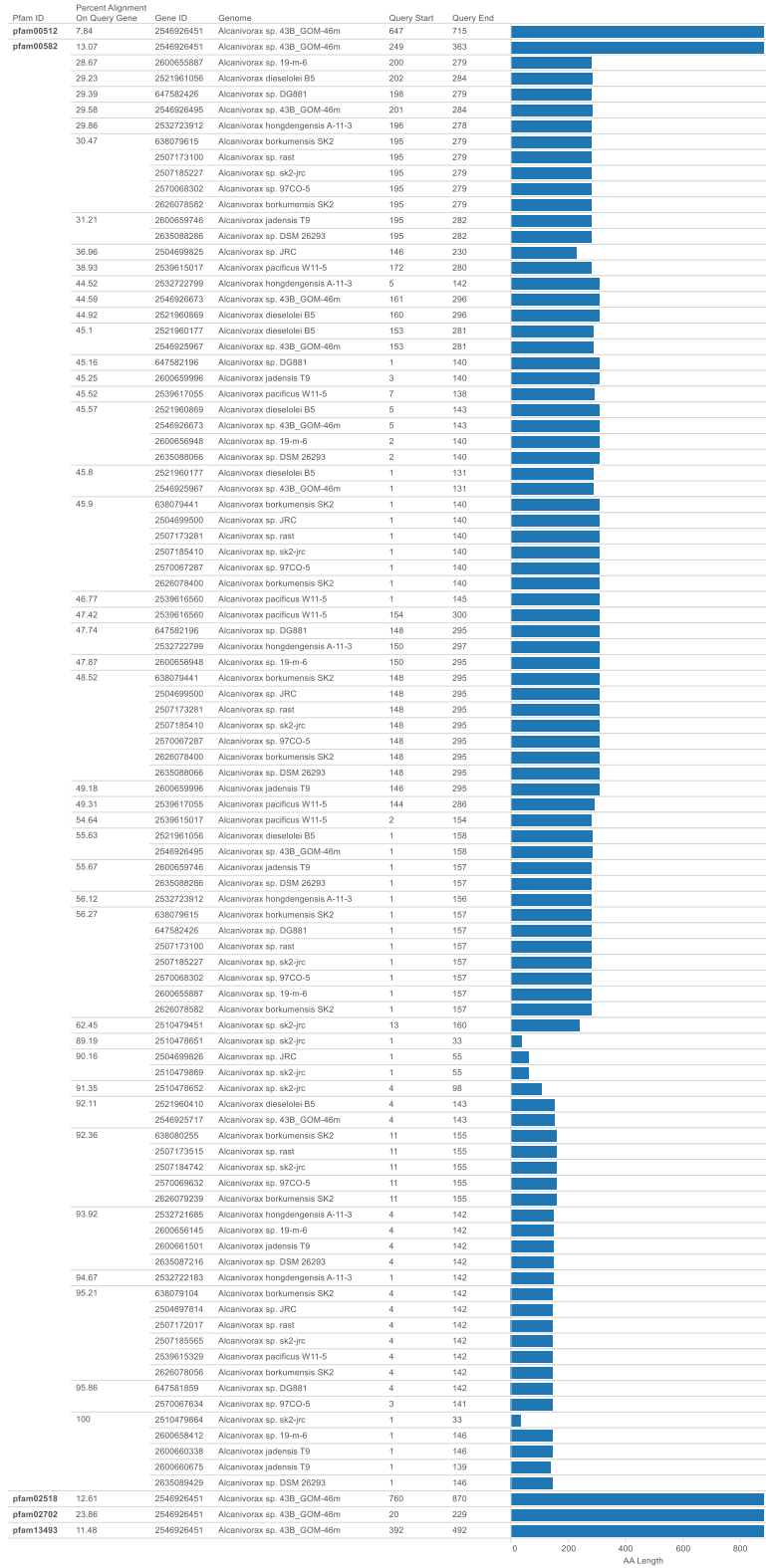


Figure 18. Protein domain organisation of *Alcanivorax* universal stress proteins.

4.2 Objective 2: To evaluate the genomic context of genes encoding the universal stress protein

4.2.1 Relative Transcription Direction of *Alcanivorax* Universal Stress Protein Genes

A total of eight *Alcanivorax* genomes were compared by the frequency of the four binary gene adjacency codes (111, 110, 011 and 010) (Figure 19). The general trend (with exception of *Alcanivorax* sp. *DG881*) was that code 111 (three genes have same transcription direction) had the highest code frequency in the genomes.

A set of 19 USP genes were grouped by the gene adjacency codes and integrated with data on protein sequence length (Figure 20). The view presents a decision support for distinguishing USPs on basis of the gene adjacency code and count of protein domain. The typical length of the USP domain is 140 aa. Thus, genes with at least double the USP domain length could be *Alcanivorax* proteins with two USP domains.

The four USP genes in the genome of the type species of *Alcanivorax*, *A. borkumensis* SK2, were grouped in code 010 (ABO_1011); 011 (ABO_1511, ABO_2141); and 111 (ABO_1340). The amino acid length for ABO_1011, ABO_1511, ABO_2141 and ABO_1340 are respectively 146 aa, 279 aa, 157aa and 305 aa. In the case of *Alcanivorax dieselolei* B5 (a proline lipid surfactant producer), the four USP genes grouped into 010 (B5T_01588 [286 aa], B5T_01819 [152 aa]) and 011 (B5T_02274 [305 aa], B5T_02456 [284 aa]) gene adjacency groups. These observations helped focus the chromosomal cassette alignment search on the USP genes of *A. borkumensis* SK2 and *A. dieselolei* B5.

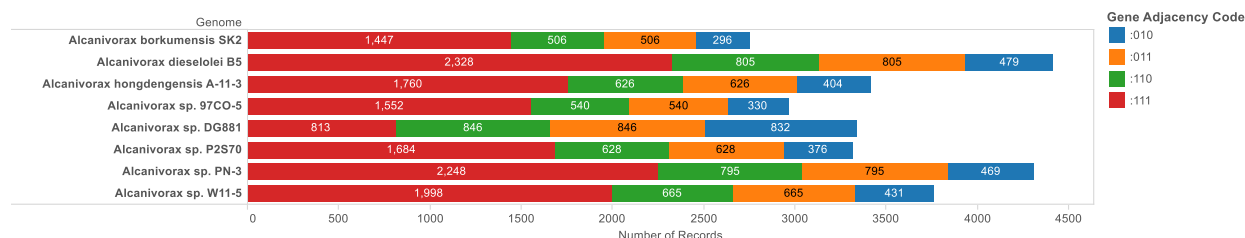


Figure 19. Relative transcription direction of adjacent genes in *Alcanivorax* genomes.

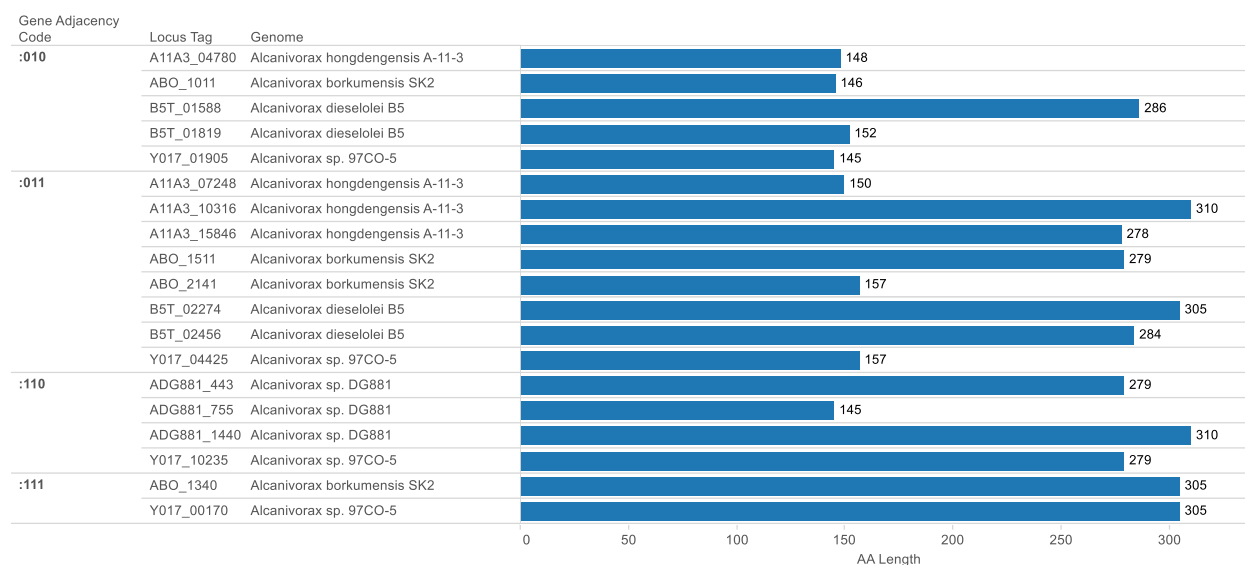


Figure 20. Gene adjacency profile and protein length of *Alcanivorax* universal stress proteins.

4.2.2 Chromosomal Cassette Search

The results of the alignment of chromosomal cassettes containing 6 *Alcanivorax* USPs are summarized in Table 9. The chromosomal cassette with ABO_1511, that has an adjacency code of 011 and encodes a 279 aa USP, aligned with genomes of *Pseudomonas* species including *Pseudomonas stutzeri*, *Pseudomonas azotifigens* and *Pseudomonas aeruginosa* (Figure 21).

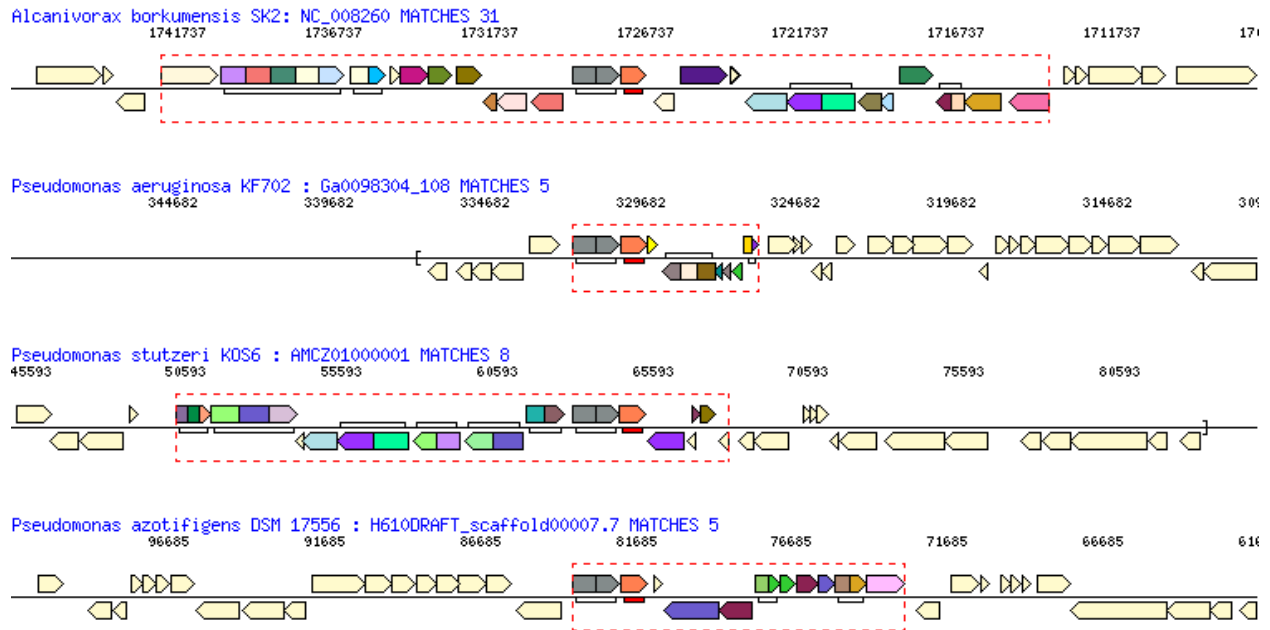


Figure 21. Example of chromosomal cassettes from non-*Alcanivorax* genomes with alignment to chromosomal cassette of *Alcanivorax borkumensis* SK2 containing universal stress protein ABO_1511.

Link to chromosomal cassette search for ABO_1511: https://img.jgi.doe.gov/cgi-bin/edu/main.cgi?section=GeneCassette&page=geneCassette&gene_oid=638079615&type=pfam

Table 9. Chromosomal cassette alignment search results for selected *Alcanivorax* genes that encode universal stress proteins.

LOCUS TAG	ALCANIVORAX USP	IMG GENE ID	GENOME SOURCE OF OTHER USP	IMG GENE OF NON-ALCANIVORAX USP	LOCUS TAG OF NON-ALCANIVORAX USP			
ABO_1011	<i>Alcanivorax borkumensis</i> SK2	638079104	<i>Haliea salexigens</i> DSM 19537	2523779912	G533DRAFT_01573			
			<i>Marine gamma proteobacterium</i> sp. HTCC2148	647649211	GPB2148_3490			
			<i>Spongiibacter tropicus</i> DSM 19543	2525572500	G411DRAFT_3062			
			<i>Gamma proteobacterium</i> BDW918	2530655760	DOK_04372			
ABO_1340	<i>Alcanivorax borkumensis</i> SK2	638079441	<i>Haliea rubra</i> CM41 15a, DSM 19751	2558343872				
			<i>Catenovulum agarivorans</i> Ymol	2550834800	WSCDRAFT_00536			
			<i>Pseudoalteromonas denitrificans</i> DSM 6059	2599841169	Ga0056025_03357			
			<i>Idimarina loihiensis</i> L2TR	637605844	IL1305			
			<i>Solimonas flava</i> DSM 18980	2527005418	K343DRAFT_00371			
ABO_1511	<i>Alcanivorax borkumensis</i> SK2	638079615	<i>Nevskia soli</i> DSM 19509	2566051656	BP26DRAFT_03902			
			<i>Pseudomonas stutzeri</i> Lautrop AB201 ATCC 17588	651018335	PSTAB_0219			
			<i>Pseudomonas azotifigens</i> DSM 17556	2523784032	H610DRAFT_01677			
			<i>Pseudomonas stutzeri</i> TS44	2532516146	Y05_00175			
			<i>Pseudomonas aeruginosa</i> PA38182	2586146055	BN889_06674			
			<i>Pseudomonas stutzeri</i> CMT. A. 9 DSM 4166	651176596	PSTAA_2725			
			<i>Pseudomonas aeruginosa</i> Ps Vir01 (initial Paired-End assembly)	2507393985	Pae-Ps Vir01_00056440			
ABO_2141	<i>Alcanivorax borkumensis</i> SK2	638080255	<i>Halopiger xanaduensis</i> SH-6	2506692353	Halxa_0545			
			<i>Deinococcus peraridilitoris</i> KR-200 DSM 19664	2509594613	Deipe_3626			
			<i>Marinobacter nanhaiticus</i> D15-8W	2546366282	J057_15045			
			<i>Metallosphaera cuprina</i> Ar-4	650821554	Mcup_1164			
			<i>Pseudanabaena</i> sp. PCC 6802	2507088070	Pse6802_2732			
			<i>Burkholderia pyrrocinia</i> Lyc2	2598941167	Ga0059242_02979			
			<i>Citricoccus</i> sp. CH26A	2548607044	CITRIDRAFT_02933			
			<i>Sulfolobus acidocaldarius</i> N8	2524414827	SacN8_07960			
			<i>Metallosphaera sedula</i> DSM 5348	640506916	Msed_1017			
			<i>Gloeobacter violaceus</i> PCC 7421	637460091	g112689			
			<i>Pseudanabaena</i> sp. PCC 7429	2504581711				
			BST_01588	<i>Alcanivorax dieselolei</i> B5	2521960177	<i>Thioalkalivibrio thiocyanodenitrificans</i> ARhDI	251623588	ThithiDRAFT_1041
						<i>Thioalkalivibrio</i> sp.ALgr5	2518633260	D893DRAFT_02343
<i>Marinobacter santoriniensis</i> NKSG1	2546447972	MSNKSG2_03245						
<i>Thiohalomonas denitrificans</i> HLD2	2595193337	Na61DRAFT_00982						
<i>Alkalilimnicola ehrlichii</i> MLHE-1	638127628	M1g_2866						
BST_01819	<i>Alcanivorax dieselolei</i> B5	2521960410	<i>Haliea salexigens</i> DSM 19537	2523779912	G533DRAFT_01573			
			<i>Marine gamma proteobacterium</i> sp. HTCC 2148	647649211	GPB2148_3490			
			<i>Melitea salexigens</i> DSM 19753	2523735261	H573DRAFT_00152			
			<i>Spongiibacter tropicus</i> DSM 19543	2525572500	G411DRAFT_3062			
			<i>Gamma proteobacterium</i> BDW918	2530655760	DOK_04372			

4.3 Objective 3: To predict biological processes involving *Alcanivorax* universal stress proteins

4.3.1 Evaluation of Stress Response Equipped Transcription Units

The Pfam functions encoded by the gene after a USP gene was used to predict stress responsive biological processes for five *Alcanivorax* USPs (Table 10). Overall, three groups of functions were in the same transcription unit (operon) with a USP gene. The gene for a 350 aa USP ABO_1340 has gene adjacency code of 111 indicating the gene before and gene after are in the same transcriptional direction. The remaining four genes (ABO_1511 & ABO_2141 from *A. borkumensis* SK12; B5T_02274 & B5T_024576 from *Alcanivorax dieselolei* B5) have the 011 gene adjacency code. The gene before ABO_1340 had hypothetical protein annotation without a specific Pfam annotation. The gene after ABO_1340 was annotated as encoding two domains: pfam00027 (Cyclic nucleotide-binding domain) and pfam13545 (Crp-like helix-turn-helix domain).

Table 10. Functions of genes adjacent to genes for *Alcanivorax* universal stress proteins

Locus Tag*	Gene Adjacency Code	Pfam after gene	Pfam name after gene
ABO_1340	111	pfam00027 - cNMP_binding	Cyclic nucleotide-binding domain
		pfam13545 - HTH_Crp_2	Crp-like helix-turn-helix domain
ABO_1511	011	pfam00916 - Sulfate_transp	Sulfate permease family
		pfam01740 – STAS	Sulphate transporter and antisigma factor antagonist (STAS)
ABO_2141	011	pfam00474 – SSF	Sodium:solute symporter family
B5T_02274	011	pfam00027 - cNMP_binding	Cyclic nucleotide-binding domain
		pfam00325 – Crp	Bacterial regulatory proteins, crp family
B5T_02456	011	pfam00916 - Sulfate_transp	Sulfate permease family
		pfam01740 - STAS	Sulphate transporter and antisigma factor antagonist (STAS)

The gene for universal stress protein ABO_1340 is part of a four-member transcription unit starting with ABO_1339 (annotated as hypothetical protein). The two genes (ABO_1341 and ABO_1342) to the right of ABO_1340 were annotated respectively as transcriptional regulator (Anr) and coproporphyrinogen III oxidase (hemF) (Figure 22). The equivalent USP gene (B5T_02274) in *Alcanivorax dieselolei* B5 has two genes to the right that are identical to those for ABO_1340 in *Alcanivorax borkumensis* SK2. B5T_022754 is predicted to be in a transcription unit with only the transcriptional regulator (Figure 23).

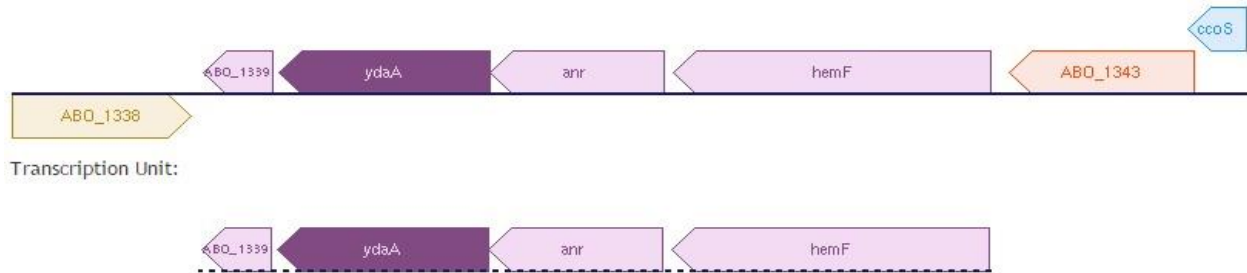


Figure 22. Genomic context and transcription unit in *Alcanivorax borkumensis* SK2 universal stress protein ABO_1340 (ydaA).



Figure 23. Genomic context and transcription unit in *Alcanivorax dieselolei* B5 universal stress protein B5T_02274 (ydaA).

Genes with locus tags B5T_02456 and ABO_1511 have an adjacent gene with function for sulphate transport. A multi-genome alignment constructed with BioCyc Multi-Genome Alignment tool for two *Alcanivorax* genomes, one *Chromohalobacter* genome and one *Hahella* genome reveal the occurrence of sulphate transport and universal stress protein a transcription unit (Figure 24).

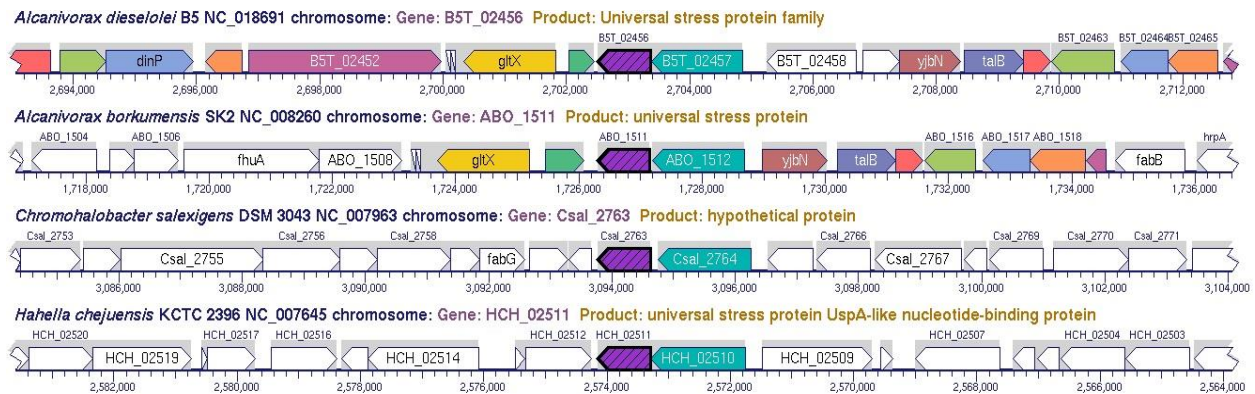


Figure 24. Multi-genome alignment of neighbourhood of genes for universal stress protein of *Alcanivorax dieselolei* and selected genomes.

4.3.2 Predicted Functional Associations of *Alcanivorax* Universal Stress Proteins

The predicted functional associations of the four *Alcanivorax borkumensis* USPs (ABO_1340, ABO_2141, ABO_1511 and ABO_1011) were determined using the Search Tool for the Retrieval of Interacting Genes/Proteins (STRING <http://string-db.org/>) (Figure 25 and Figure 26). ABO_1340 had the highest association with 10 nodes, while others had one (ABO_2141), two (ABO_1011) to three (ABO_1511) nodes (Figure 25). The 10 predicted functional partners for *Alcanivorax borkumensis* ABO_1340 were transcriptional regulator Anr (ABO_1341), hypothetical proteins (ABO_1339, ABO_1346, ABO_0534, ABO_1043, ABO_0686 and ABO_1044) coproporphyrinogen III oxidase (ABO_1342), chaperonin (ABO_0634) and cation-transporting P-type ATPase (ABO_1345). The three partners of ABO_1340 had high combined score [>0.7], a measure of confidence of the interaction, were

ABO_1339, ABO_1341 and ABO_1342. ABO_1011 had two predicted functional partners with medium combined confidence score: chaperonin (ABO_0634) and DNA topoisomerase I (ABO_1012). The functional partners for ABO_1511 were sulfate permease (ABO_1512) and sulfate transporters (ABO_1943, ABO_1945). The interaction between ABO_1511 and ABO_1512 has a high combined score (0.973) with evidence from gene neighbourhood and co-expression. In the case of ABO_2141 the functional partner was a sodium solute transporter family protein (ABO_2142) with a high combined confidence score of 0.859.

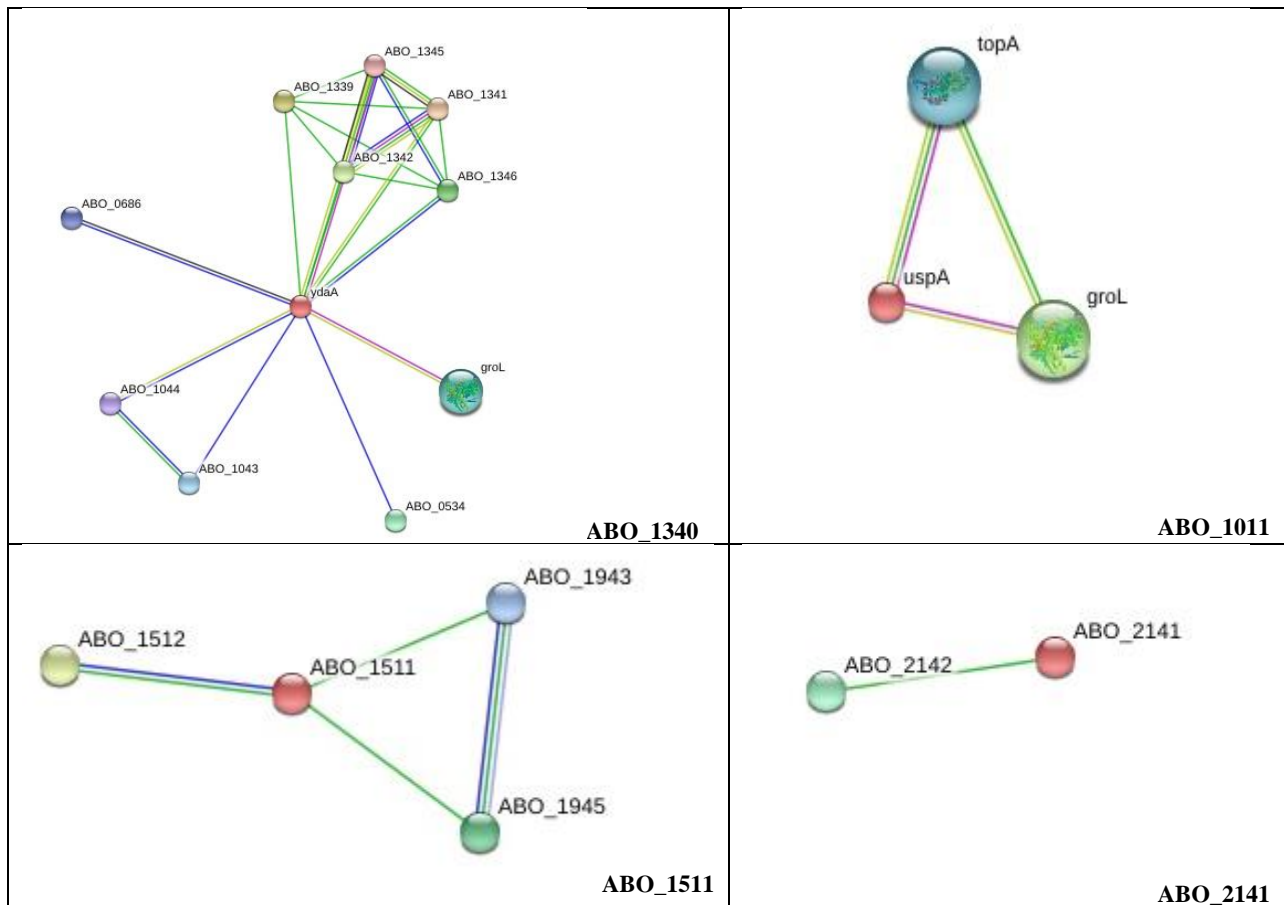


Figure 25. Interaction views from *Alcanivorax borkumensis* locus tags ABO_1340, ABO_2141, ABO_1511 and ABO_1011

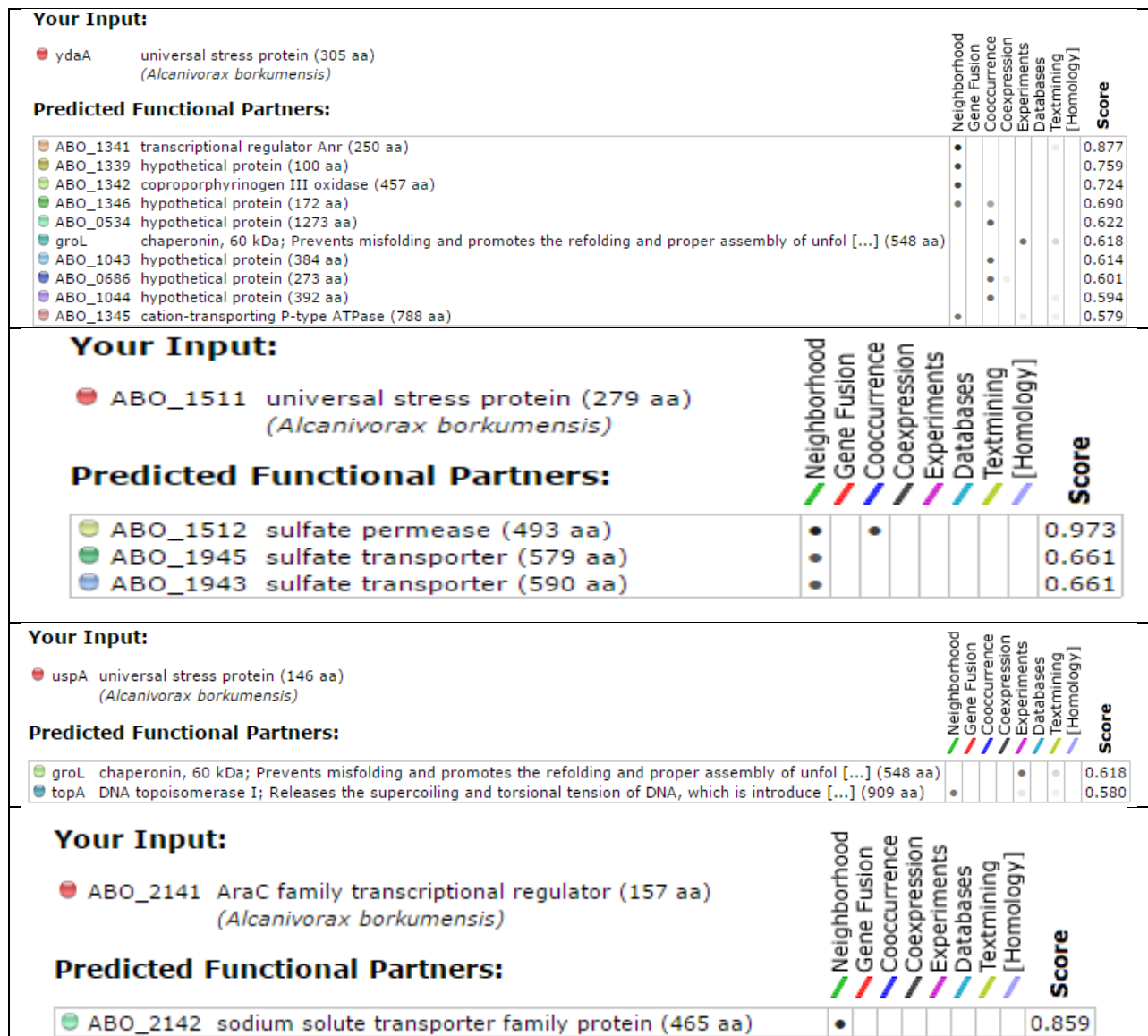


Figure 26. Confidence scores for interactions with functional partners to *Alcanivorax borkumensis* universal stress protein

4.3.3 Expression Levels of Genes in Transcription Units

The NCBI Gene Expression Omnibus (<http://www.ncbi.nlm.nih.gov/geo>) under accession number GSE44687 [*A. borkumensis* cells: control vs. 1-octanol-stressed cells] (Naether *et al.* 2013) was selected for analysis of expression levels of *Alcanivorax borkumensis* SK2 genes encoding universal stress proteins and their partners in transcription units. The primary reason was to focus on stress associated with 1-octanol, the most toxic compound accumulating as

degradation intermediate in bacteria grown with an n-alkane as the substrate (Naether *et al.* 2013). Additionally, the gene expression values have positive and negative values which enable comparative analysis for patterns. For example, genes in same transcription units are expected to have shared the same direction of gene expression.

The samples in the gene expression series GSE44687 (<http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE44687>;) were generated using the Agilent microarray platform (GPL16725). The Agilent-026176 *Alcanivorax borkumensis* SK2 can be accessed at <http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GPL16725>. The accession numbers of the 31 samples were from GSM1088974 to GSM1089004 (Table 11). The gene expression series is based on *Alcanivorax borkumensis* cells (grown on either pyruvate or hexadecane as carbon source) that were stressed with 1- octanol and data collected at 15 min, 30 min, 60 min and 90 min after 1-octanol addition. The microarray platform has three probes per gene (locus tag) (Table 12).

Table 11. Sample information for GSE44687 gene expression profiling of *Alcanivorax borkumensis* cells: control vs. 1-octanol-stressed cells

!Sample title	Sample	!Sample Group
CellsHex_0min_noOctanol_rep1	GSM1088990	CellsHex_0min_noOctanol
CellsHex_0min_noOctanol_rep2	GSM1088993	CellsHex_0min_noOctanol
CellsHex_0min_noOctanol_rep3	GSM1089001	CellsHex_0min_noOctanol
CellsHex_15 min_1.64mMOctanol_rep1	GSM1088991	CellsHex_15 min_1.64mMOctanol
CellsHex_15 min_1.64mMOctanol_rep2	GSM1088992	CellsHex_15 min_1.64mMOctanol
CellsHex_15 min_1.64mMOctanol_rep3	GSM1088994	CellsHex_15 min_1.64mMOctanol
CellsHex_30 min_1.64mMOctanol_rep1	GSM1088995	CellsHex_30 min_1.64mMOctanol
CellsHex_30 min_1.64mMOctanol_rep2	GSM1088996	CellsHex_30 min_1.64mMOctanol
CellsHex_30 min_1.64mMOctanol_rep3	GSM1088997	CellsHex_30 min_1.64mMOctanol
CellsHex_60 min_1.64mMOctanol_rep1	GSM1088998	CellsHex_60 min_1.64mMOctanol
CellsHex_60 min_1.64mMOctanol_rep2	GSM1088999	CellsHex_60 min_1.64mMOctanol
CellsHex_60 min_1.64mMOctanol_rep3	GSM1089000	CellsHex_60 min_1.64mMOctanol
CellsHex_90 min_1.64mMOctanol_rep1	GSM1089002	CellsHex_90 min_1.64mMOctanol
CellsHex_90 min_1.64mMOctanol_rep2	GSM1089003	CellsHex_90 min_1.64mMOctanol
CellsHex_90 min_1.64mMOctanol_rep3	GSM1089004	CellsHex_90 min_1.64mMOctanol
CellsPyr_0min_noOctanol_rep1	GSM1088974	CellsPyr_0min_noOctanol
CellsPyr_0min_noOctanol_rep2	GSM1088978	CellsPyr_0min_noOctanol
CellsPyr_0min_noOctanol_rep3	GSM1088982	CellsPyr_0min_noOctanol
CellsPyr_0min_noOctanol_rep4	GSM1088986	CellsPyr_0min_noOctanol
CellsPyr_15min_0.66mMOctanol_rep1	GSM1088975	CellsPyr_15min_0.66mMOctanol
CellsPyr_15min_0.66mMOctanol_rep2	GSM1088976	CellsPyr_15min_0.66mMOctanol
CellsPyr_15min_0.66mMOctanol_rep3	GSM1088977	CellsPyr_15min_0.66mMOctanol
CellsPyr_30min_0.66mMOctanol_rep1	GSM1088979	CellsPyr_30min_0.66mMOctanol
CellsPyr_30min_0.66mMOctanol_rep2	GSM1088980	CellsPyr_30min_0.66mMOctanol
CellsPyr_30min_0.66mMOctanol_rep3	GSM1088981	CellsPyr_30min_0.66mMOctanol
CellsPyr_60min_0.66mMOctanol_rep1	GSM1088983	CellsPyr_60min_0.66mMOctanol
CellsPyr_60min_0.66mMOctanol_rep2	GSM1088984	CellsPyr_60min_0.66mMOctanol
CellsPyr_60min_0.66mMOctanol_rep3	GSM1088985	CellsPyr_60min_0.66mMOctanol
CellsPyr_90min_0.66mMOctanol_rep1	GSM1088987	CellsPyr_90min_0.66mMOctanol
CellsPyr_90min_0.66mMOctanol_rep2	GSM1088988	CellsPyr_90min_0.66mMOctanol
CellsPyr_90min_0.66mMOctanol_rep3	GSM1088989	CellsPyr_90min_0.66mMOctanol

Table 12. Probe identifiers for genes for universal stress proteins and adjacent genes in the Agilent-026176 *Alcanivorax borkumensis* SK2 microarray platform (GPL16725)

Probe ID	Locus Tag	Description
CUST_3028_PI420440664	ABO_1010	hypothetical protein ABO_1010
CUST_3029_PI420440664	ABO_1010	hypothetical protein ABO_1010
CUST_3030_PI420440664	ABO_1010	hypothetical protein ABO_1010
CUST_3031_PI420440664	ABO_1011	universal stress protein
CUST_3032_PI420440664	ABO_1011	universal stress protein
CUST_3033_PI420440664	ABO_1011	universal stress protein
CUST_3034_PI420440664	ABO_1012	DNA topoisomerase I
CUST_3035_PI420440664	ABO_1012	DNA topoisomerase I
CUST_3036_PI420440664	ABO_1012	DNA topoisomerase I
CUST_4015_PI420440664	ABO_1339	hypothetical protein ABO_1339
CUST_4016_PI420440664	ABO_1339	hypothetical protein ABO_1339
CUST_4017_PI420440664	ABO_1339	hypothetical protein ABO_1339
CUST_4018_PI420440664	ABO_1340	universal stress protein
CUST_4019_PI420440664	ABO_1340	universal stress protein
CUST_4020_PI420440664	ABO_1340	universal stress protein
CUST_4021_PI420440664	ABO_1341	transcriptional regulator Anr
CUST_4022_PI420440664	ABO_1341	transcriptional regulator Anr
CUST_4023_PI420440664	ABO_1341	transcriptional regulator Anr
CUST_4528_PI420440664	ABO_1510	hypothetical protein ABO_1510
CUST_4529_PI420440664	ABO_1510	hypothetical protein ABO_1510
CUST_4530_PI420440664	ABO_1510	hypothetical protein ABO_1510
CUST_4531_PI420440664	ABO_1511	universal stress protein
CUST_4532_PI420440664	ABO_1511	universal stress protein
CUST_4533_PI420440664	ABO_1511	universal stress protein
CUST_4534_PI420440664	ABO_1512	sulfate permease protein
CUST_4535_PI420440664	ABO_1512	sulfate permease protein
CUST_4536_PI420440664	ABO_1512	sulfate permease protein
CUST_6418_PI420440664	ABO_2140	hypothetical protein ABO_2140
CUST_6419_PI420440664	ABO_2140	hypothetical protein ABO_2140
CUST_6420_PI420440664	ABO_2140	hypothetical protein ABO_2140
CUST_6421_PI420440664	ABO_2141	AraC family transcriptional regulator
CUST_6422_PI420440664	ABO_2141	AraC family transcriptional regulator
CUST_6423_PI420440664	ABO_2141	AraC family transcriptional regulator
CUST_6424_PI420440664	ABO_2142	sodium solute transporter family protein
CUST_6425_PI420440664	ABO_2142	sodium solute transporter family protein
CUST_6426_PI420440664	ABO_2142	sodium solute transporter family protein

Genes encoding universal stress protein domain (pfam00582) highlighted in bold.

The average expression level across all the samples for each universal stress protein and the adjacent genes was calculated to provide an overview of the pattern of average expression level for the 31 samples (Figure 27). The adjacent genes of the USP genes in the same direction of transcription had also the same direction of expression. ABO_1340 (USP) and ABO_1341 (transcriptional regulator Anr) both have positive gene expression values. The following gene pairs have negative expression values: (i) ABO_1511 (universal stress protein) and ABO_1512 (sulfate permease protein); (ii) ABO_2141 (AraC family transcriptional regulator) and ABO_2142 (sodium solute transporter family protein).

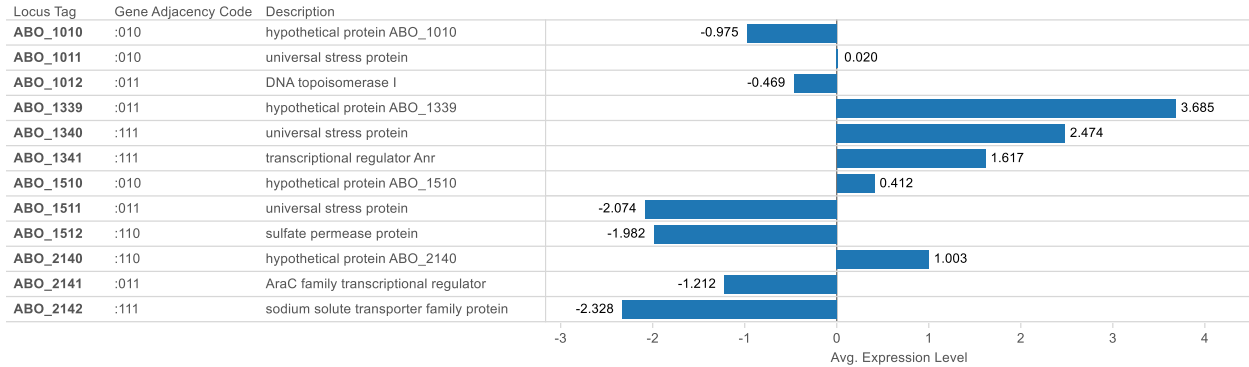


Figure 27. Average expression levels of *Alcanivorax borkumensis* genes for universal stress protein and adjacent genes.

Average expression levels was calculated from expression levels of the 31 samples in the GSE44687 series GSM1088974 to GSM1089004 (Table 11).

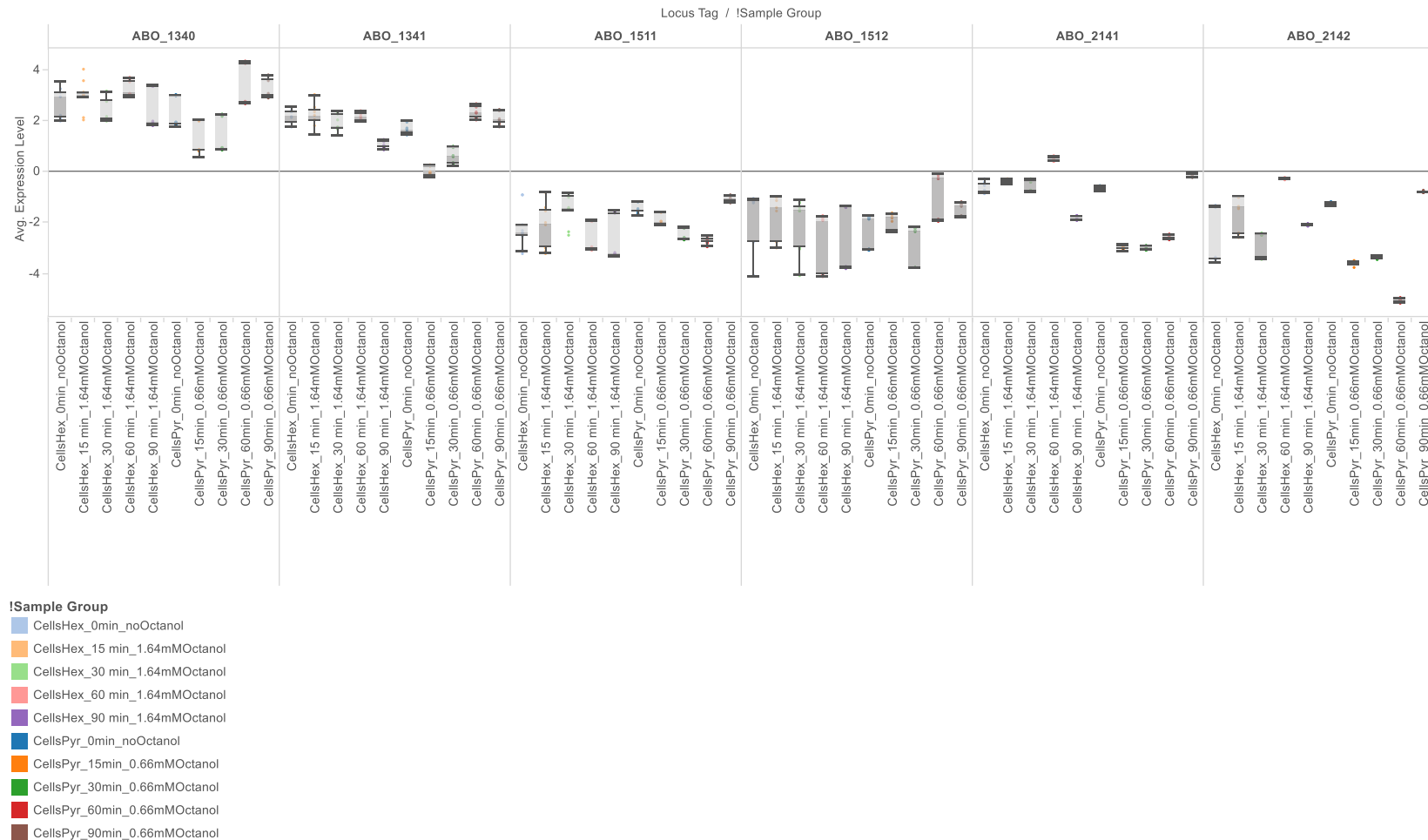


Figure 28. Overview of expression levels of *Alcanivorax borkumensis* probes for genes annotated for universal stress protein domain and their adjacent genes in same transcription direction.

Box Plot display distribution of expression values for probes on microarray slide that targets universal stress protein and adjacent genes in same transcription direction. Data obtained from Naether et al. (2013). The gene expression series is based on *Alcanivorax borkumensis* cells (grown on either pyruvate or hexadecane as carbon source) that were stressed with 1-octanol and data collected at 15 min, 30 min, 60 min and 90 min after 1-octanol addition.

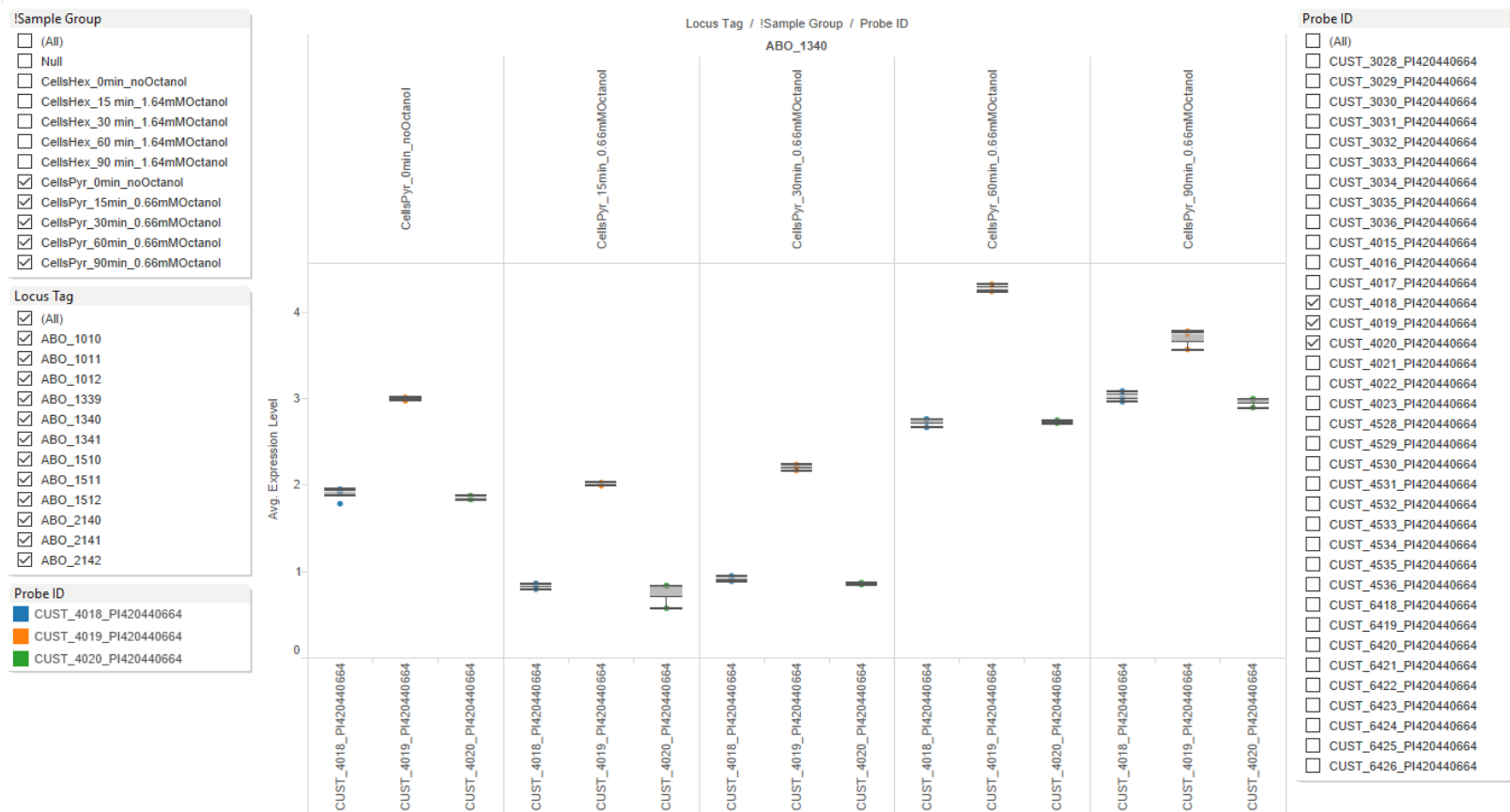


Figure 29. Interface for discovering expression levels of probes of *Alcanivorax borkumensis* for universal stress protein and adjacent genes.

Dashboard system for discovering relationships between expression levels of microarray probes and stress (concentration and duration of exposure). The Box Plot reveals the distribution of the expression levels obtained in the three samples (replicates) in the sample group (Sample Group). Data obtained from Naether et al. (2013). The gene expression series is based on *Alcanivorax borkumensis* cells (grown on either pyruvate or hexadecane as carbon source) that were stressed with 1- octanol and data collected at 15 min, 30 min, 60 min and 90 min after 1-octanol addition.

CHAPTER 5

DISCUSSION AND CONCLUSIONS

Alcanivorax borkumensis SK2, the most studied species among the marine hydrocarbonoclastic bacteria, has so far been cultivated only on different hydrocarbons and pyruvate as sole carbon and energy sources (Yakimov *et al.*, 2007). Due to its ability to utilize a broad variety of aliphatic hydrocarbons, *A. borkumensis* blooms during oil spills and can represent up to 80 to 90% of the associated bacterial community (Hara *et al.*, 2004; Yakimov *et al.*, 2005). Because of its outstanding importance to bioremediation of oil-contaminated marine environments, its genome has been sequenced (Schneiker *et al.*, 2006), and proteome studies have also been performed (Sabirova *et al.*, 2006).

The use of bioinformatics approaches to available genomics data present abundant opportunities to predict and understand the biological functions of *Alcanivorax* species. This understanding can lead to solutions to improve, or harness the characteristic of *Alcanivorax* species which can be beneficial industrially or environmentally. A functional insight into stress response of *Alcanivorax* could help to improve the adaption system of the organism in oil spilled environment and thus use them in bioremediation processes.

The dissertation research is the first to investigate the stress response systems that involved genes for the *Alcanivorax* universal stress proteins (USPs). The research investigation also utilized visual analytics techniques to represent, analyze and interact with results obtained from bioinformatics software. The application of visual analytics techniques available through visual analytics software has provided deeper insights on results obtained from the three research objectives. This section discusses the results for each objective and provides an integrated

discussion of key findings. The dissertation concludes with a conclusions and suggestions for future research.

5.1 Objective 1: To make comparison among the protein domain organization of universal stress proteins encoded in *Alcanivorax* genomes.

5.1.1 Genome Statistics of *Alcanivorax* in the Integrated Microbial Genomes Database

A total of 15 *Alcanivorax* genomes were retrieved from the Integrated Microbial Genome (Table 7). Thus, this dissertation research has extracted comprehensive information on *Alcanivorax* genomes. The percentage Guanosine+Cytosine (GC%) content of the *Alcanivorax* genome ranged from 55% to 63%, with the type species *Alcanivorax borkumensis* SK2 with a GC percentage of 55%. The variations in GC content may reflect habitat of the *Alcanivorax* species (Hildebrand *et al.*, 2010). Apart from the two *Alcanivorax borkumensis* SK2 genomes, five other genomes had GC% of 55%. On the basis of identical GC content, these seven genomes could be derived from strains of *Alcanivorax borkumensis*. Indeed three genomes with IMG Identifiers (Genome Names in brackets) 637000004 (*Alcanivorax borkumensis* SK2), 2507149004 (*Alcanivorax* sp. Rast) and 250749008 (*Alcanivorax* sp. Sk2-jrc) are grouped in the same cluster by a measure of genome similarity termed Whole-genome based Average Nucleotide Identity (gANI) [<https://ani.jgi-psf.org/html/individualClusters.php?clusterId=783>] (Varghese *et al.*, 2015). Comparative analysis of the gene annotations in these similar *A. borkumensis* genomes could help identify common genes and strain-specific genes.

The GC% *Alcanivorax borkumensis* SK2 (55%) is higher for the following *Oceanospirillales* taxa: *Bermanella marisrubri* RED65 (44%); *Gynuella sunshinyii* YC6258 (49%); *Kangiella koreensis* SW-125, DSM 16069 (44%); *Marinomonas mediterranea* MMB-1, ATCC 700492 (44%); *Marinomonas* sp. MWYL1 (43%); *Neptuniibacter caesariensis* MED92

(47%); and *Oleispira antarctica* RD-8 (42%) (Kube *et al.* 2013; Markowitz *et al.* 2012). Compared to *A. borkumensis* SK2 genomes, *Hahella chejuensis* KCTC 2396 (54%) had similar GC content of 54% while *Chromohalobacter salexigens* 1H11, DSM 3043 had a higher GC% of 64%. The genome of *A. borkumensis* SK2 has chromosomal regions with GC% that significantly differ from 55% and may reflect regions of horizontal gene transfer (Reva *et al.*, 2008). Studies by Wu *et al.*, (2012), have shown that DNA polymerase III α subunit and its isoform which take part either in replication or SOS mutagenesis play a key role in determining GC variability. Also environmental or bacteriological factors i.e. temperature, habitat, genome size, oxygen requirement play vital roles or depend indirectly on various mutator genes to get the actual GC content (Wu *et al.*, 2012). Therefore GC content is crucial in organisms adapting to an environment either during stressed or normal conditions.

5.1.2 Gene count and protein domain organization of universal stress proteins encoded in *Alcanivorax* genomes.

The genes per genome encoding the universal stress protein domain ranged from three to five in the 15 *Alcanivorax* genomes investigated. Eleven genomes had four USP genes with the type species and another finished genome (*Alcanivorax dieselolei* B5) having four genes (Table 8). This USP gene count is lower than reported in *Escherichia coli* K-12, MG1655 (7 USP genes) (Nachin *et al.*, 2005) and *Mycobacterium tuberculosis* H37Rv (10 USP genes) (Hingley-Wilson *et al.*, 2010). The locus tags for the four *Alcanivorax borkumensis* SK2 genes annotated with the universal stress protein domain are ABO_1011, ABO_1340, ABO_1511 and ABO_2141. Additional sequence and database evaluations reveal that ABO_1340 and ABO_1511 have comprehensive support as universal stress proteins.

The genome of *A. dieselolei* B5 encoded three USPs (B5T_01588, B5T_02274 and B5T_02456) with double USP domains while *A. borkumensis* SK2 encoded two USPs (ABO_1340 and ABO_1511) with double USP domains (Figure 16). In *Escherichia coli* UspE (*ydaA*) has two USP units E1 and E2 and they function in oxidative stress defence as well as promote the aggregation of cells (Nachin *et al.*, 2005). Aggregation of cells is an initial step for formation of biofilms (Claessen *et al.*, 2014). The function of the *Alcanivorax* USPs has not been investigated for biochemical and phenotypic functions. An objective for further research could be to determine the function of *Alcanivorax* USPs with two USP units in the aggregation of cells and subsequent biofilm formation. Biofilm formation is a stress response strategy used by the aerobic *A. borkumensis* to cope with environmental challenges in marine and coastal habitats (Sabirova *et al.*, 2008). Further, biofilm formation is for faster and intense depletion of linear and branched hydrocarbon (Dasgupta *et al.*, 2013; Gertler *et al.*, 2009).

The protein domain organization annotation for the 62 *Alcanivorax* USP genes showed that in addition to the USP domain (pfam00582), gene 2546926451 (Locus Tag: N537DRAFT_02674) of *Alcanivorax sp.43_GOM-46m* had annotations for pfam00512 [His Kinase A (phospho-acceptor) domain], pfam02518 [Histidine kinase-, DNA gyrase B-, and HSP90-like ATPase], pfam02702 (Osmosensitive K⁺ channel His kinase sensor domain) and pfam13493 [Domain of unknown function (DUF4118)]. The focus of this research is on proteins that contain only the USP domain. ABO_2141 was observed in only *Alcanivorax borkumensis* genomes (Figure 18). This raises need for additional analysis of the sequence to clarify if ABO_2141 encodes a universal stress protein. A total of 12 functional (ligand) binding sites characterize universal stress proteins and includes an ATP-binding motif G-2X-G-9X-G(S/T) (Arembinski, Ung, Oachim, Ieckmann, & Im, 1998; Isokpehi, Simmons, et al., 2011). The NCBI

Conserved Domain Search web page for WP_011589419.1 (ABO_2141) provides the identical functional sites to the UspA domain of MJ0577 (also called 1MJH) from *Methanocaldococcus jannaschii*. Only three ligand binding sites (glycine, serine and valine) were observed for ABO_2141. This is in contrast with presence the complete ATP-binding motif in the two *Alcanivorax* protein sequences with accession WP_011588622.1 (ABO_1340) and WP_011588792.1 (ABO_1511).

5.2 Objective 2: To evaluate the genomic context of genes encoding the universal stress protein

5.2.1 Relative Transcription Direction of *Alcanivorax* Universal Stress Protein Genes

The gene adjacency profiles of eight *Alcanivorax* genomes revealed that 1447 genes had gene neighbours in the same direction (gene adjacency code 111) for the genome of *Alcanivorax borkumensis* SK2 (Figure 19). The comparative visual representation provides an overview of the number of genes that could be in operons where the adjacent genes are in the same direction as the gene. The gene adjacency code profiling of eight genomes revealed that only *Alcanivorax dieselolei* B5 and *Alcanivorax sp.* PN-3 had gene count greater than 4,000 (Figure 19). *Alcanivorax sp.* PN-3 was isolated from beach sands contaminated with crude oil discharged from the Deepwater Horizon spill (Overholt *et al.*, 2013). The protein-coding gene count reported for two another strain of *A. dieselolei* was 4445 genes for strain KS-293 from surface seawater in the eastern Mediterranean Sea (Barbato *et al.*, 2015). The functions of the additional genes of these genomes compared to *Alcanivorax borkumensis* SK2, which has protein-coding gene count of 2755, warrants further research. The genomic context revealed the universal stress proteins found in *Alcanivorax*, this could be checked for in other potential oil degraders and they can serve as substitute or even better choices when planning an oil clean-up exercise.

The 305aa USP, ABO_1340 had a transcription unit of gene adjacency code of 111 (Figure 20). The visual representation in Figure 20 integrates the data on gene adjacency code, locus tag, genome name and protein length. Therefore, ABO_1340 and Y017_00170 have identical gene adjacency profile as well as protein (amino acid length). The protein sequences have 100% identity and identical adjacent genes *Alcanivorax sp.* strain 97CO-5 was isolated from a crude-oil-degrading consortium, enriched from Yellow Sea sediment of China (Luan *et al.*, 2014). The relative transcription direction considered with gene order has given a lot of insight into the mechanism of mutations and gene movement such as transposition that may have occurred during speciation and thus evolution, this is very crucial for studying adaptation of *Alcanivorax* to an oil spill environment.

5.2.2 Chromosomal Cassette Search

The chromosomal cassette with ABO_1511, that has an adjacency code of 011 and encodes a 279 aa USP, aligned with genomes of *Pseudomonas* species including *Pseudomonas stutzeri*, *Pseudomonas azotifigens* and *Pseudomonas aeruginosa* (Figure 21). The universal stress proteins of *P. aeruginosa* have been characterized for response to anaerobic energy stress (Boes *et al.*, 2006; Boes, Schreiber, & Schobert, 2008; Schreiber *et al.*, 2006). The bacteria taxa that have similar chromosomal content could be useful for comparative analysis of the universal stress proteins of *Alcanivorax* species. This can also help identify closely related species into which the ability to utilize long chain hydrocarbons can be introduced by transformation with one or two genes derived from *Alcanivorax*.

5.3. Objective 3: To predict biological processes involving *Alcanivorax* universal stress proteins

5.3.1 Evaluation of Stress Response Equipped Transcription Units

Among the four *Alcanivorax borkumensis* genes annotated for USP domain, ABO_1340 and ABO_1511 are in transcription units which are also found in at least one other genome. ABO_1340 is adjacent to a gene encoding cyclic nucleotide-binding domain (ABO_1341). The function of ABO_1341 is not known but proteins with cyclic nucleotide-binding domains have functions including protein phosphorylation, ion conductance and regulation of gene transcription (Shabb *et al.*, 1992). Interestingly the gene description annotation for ABO_1341 is transcriptional regulator Anr. The coproporphyrinogen II oxidase gene was also annotated as hemF and in same transcription unit with *usp* and *anr* genes (Figure 22). The hemF is an enzyme that requires oxygen to function in heme biosynthesis and catalyses the oxidative decarboxylation of coproporphyrinogen III to form protoporphyrinogen IX (Troup *et al.*, 1994; Zeilstra-ryalls, Schornberg, Hemf-up, & The, 2006). In *Rhodobacter sphaeroides* 2.4.1 *hemF* gene is required for aerobic growth and oxygen controls *hemF* expression (Zeilstra-ryalls *et al.*, 2006). The chromosomal region ABO_1340 to ABO_1351 contains genes in the same transcription unit and with the following predicted functions cation-transporting P-type ATPase; cytochrome c oxidase, cbb3-type, CcoQ subunit, putative; cytochrome c oxidase, cbb3-type, subunit I; cytochrome c oxidase, cbb3-type, subunit II; cytochrome c oxidase, cbb3-type, subunit III, putative; cytochrome oxidase maturation protein, cbb3-type; and iron-sulfur cluster-binding protein. Altogether, the genomic neighbourhood of the USP in locus tag ABO_1340 is enriched for oxygen sensing. For example, the expression of cytochrome cbb3 oxidase facilitates human pathogens to colonize tissues that lack oxygen (Pitcher and Watmough 2004). Additionally, iron-

sulfur clusters act as biological sensors of environmental conditions including oxygen and oxidative stress (Crack *et al.*, 2014).

The *Alcanivorax borkumensis* USP, ABO_1511 is adjacent to a gene encoding sulphate transporter (ABO_1512). ABO_1511 has the same adjacency code and same amino acid length of 279 aa with A11A3_15846. Sulphate is taken into the cell by specific transporters known as sulphate permease. When there is favourable organic sulphur source in the environment, the cell immediately switches off the ATP-dependent sulphate assimilation pathway beginning with sulphate permease (Piłsyk & Paszewski, 2009). Sulphate transporters have being shown to be important when there is high salinity (Cao *et al.*, 2014) and drought stress (Gallardo, Courty, Le Signor, Wipf, & Vernoud, 2014). Sulphate transporters also play a vital role in plants in uptake and distribution to other parts of the plant (Buchner, Takahashi, & Hawkesford, 2004).

The Search Tool for the Retrieval of Interacting Genes/Proteins (STRING <http://string-db.org/>) revealed that of the *Alcanivorax* USPs, ABO_1340 had the highest association with 10 nodes, while others had one (ABO_2141), two (ABO_1011) to three (ABO_1511) nodes (Figure 25). The three partners of ABO_1340 had high combined score [>0.7], a measure of confidence of the interaction, were ABO_1339, ABO_1341 (cyclic nucleotide-binding domain) and ABO_1342 (coproporphrinogen II oxidase). These proteins are encoded by genes adjacent to ABO_1340. The STRING functional network prediction for the *Alcanivorax borkumensis* locus ABO_2141 showed only one partner, ABO_2142 (pfam00474: sodium solute/substrate transporter family protein), with a high combined confidence score of 0.859 (Figure 25). The solutes transported through coupling with transport of sodium ions include sugars, proline, panthothenate and iodide (Jung, 2002). The annotation on the Integrated Microbial Genomes (IMG) gene page for ABO_2142 includes annotation for proline as the amino acid solute. In the

context of stress response in bacteria, sodium solute transporters function in cell adaptation to osmotic stress (Spiegelhalter and Bremer 1998). Since oceans have a high concentration of sodium ions, *Alcanivorax borkumensis* gene ABO_2142 could be relevant to regulating the content of sodium ions in the cell.

5.3.3 Expression Levels of Genes in Transcription Units

The gene expression series is based on *Alcanivorax borkumensis* cells (grown on either pyruvate or hexadecane as carbon source) that were stressed with 1-octanol and data collected at 15 min, 30 min, 60 min and 90 min after 1-octanol addition (Naether *et al.*, 2013). The adjacent genes of the ABO_1340 is part of a transcription unit. Genes for ABO_1339 and ABO_1341 are in same transcription direction with ABO_1340 as well as same direction of expression (Figure 27). Operon pairs (adjacent gene pairs in transcription units) have co-expression levels of adjacent gene pairs (Okuda *et al.*, 2007). Gene expression increases linearly with the distance from the start of a gene to the end of the operon (transcription distance) (Lim *et al.*, 2011). A knowledge of genes expressed helps to know what genes are present and what they can do, also the conditions when they are optimally expressed can help to set working parameters when going to actual field for real oil clean-up for effective results.

5.3.4 Conclusions, Limitations of Findings and Future Work

The combination of bioinformatics and visual analytics evaluation of genome-enabled data on *Alcanivorax* universal stress proteins has resulted in key biological observations and areas for future research. On the basis of transcription unit and adjacent genes, two types of *Alcanivorax* USPs genes observed were (i) adjacent to cyclic nucleotide-binding and oxygen sensing functions; and (ii) adjacent to sulfate transporter function. Both types of genes encode two universal stress protein domains (pfam00582) also referred to as tandem-type universal

stress proteins. This dissertation research evaluated data from *Alcanivorax borkumensis* cells (grown on either pyruvate or hexadecane as carbon source) that were stressed with 1-octanol and data collected at 15 min, 30 min, 60 min and 90 min after 1-octanol addition (Naether *et al.*, 2013). The two genes for *Alcanivorax borkumensis* SK2 universal stress proteins, ABO_1340 and ABO_1511, had the same direction of expression for adjacent genes. The probe sequences for ABO_1340 had positive gene expression levels for all the conditions. However, the probe sequences of ABO_1511 had negative gene expression levels.

A limitation of this research was that findings based on bioinformatics and visual analytics methods may need confirmation by molecular methods. The differences observed may also reflect the quality of the annotations provided for genes. The sequence and structural characteristics of each of the four USP domains in *Alcanivorax* needs to be further investigated. For example, do the *Alcanivorax* USP domain units have ATP-binding capability? Further research is needed on the relationship between number, length and order of genes in operons that include genes for universal stress proteins. The knowledge discovered from the genome context analytics could contribute to improving the performance of *Alcanivorax* in bioremediation of petroleum polluted environments.

REFERENCES

- Aas, E., Baussant, T., Balk, L., Liewenborg, B., & Andersen, O. K. (2000). PAH metabolites in bile, cytochrome P4501A and DNA adducts as environmental risk parameters for chronic oil exposure: a laboratory experiment with Atlantic cod. *Aquatic Toxicology*, *51*(2), 241–258.
- Zarembinski, T. I., Hung, L-W., Mueller-Dieckmann H-J., Kim K.-K., Yoko H., Kim R., & Kim S.-H. (1998). Structure-based assignment of the biochemical function of a hypothetical protein : A test case of structural genomics, *Proc Natl Acad Sci.* *95*(2), 15189–15193.
- Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., & others. (2000). Gene Ontology: tool for the unification of biology. *Nature genetics*, *25*(1), 25–29.
- Barbato, M., Mapelli, F., Chouaia, B., Crotti, E., Daffonchio, D., & Borin, S. (2015). Draft genome sequence of the hydrocarbon-degrading bacterium *Alcanivorax dieselolei* KS-293 isolated from surface seawater in the eastern Mediterranean Sea. *Genome announcements*, *3*(6), e01424–15.
- Boes, N., Schreiber, K., Härtig, E., Jaensch, L., & Schobert, M. (2006). The *Pseudomonas aeruginosa* universal stress protein PA4352 is essential for surviving anaerobic energy stress. *Journal of bacteriology*, *188*(18), 6529–6538.
- Boes, N., Schreiber, K., & Schobert, M. (2008). SpoT-triggered stringent response controls usp gene expression in *Pseudomonas aeruginosa*. *Journal of bacteriology*, *190*(21), 7189–7199.
- Bruns, A., & Berthe-Corti, L. (1999). *Fundibacter jadensis* gen. nov., sp. nov., a new slightly halophilic bacterium, isolated from intertidal sediment. *International journal of systematic bacteriology*, *49*(2), 441–448.
- Buchner, P., Takahashi, H., & Hawkesford, M. J. (2004). Plant sulphate transporters: co-ordination of uptake, intracellular and long-distance transport. *Journal of Experimental Botany*, *55*(404), 1765–1773.
- Cao, M.-J., Wang, Z., Zhao, Q., Mao, J.-L., Speiser, A., Wirtz, M., R. Hell, J. K. Zhu, & Xiang, C.-B. (2014). Sulfate availability affects ABA levels and germination response to ABA and salt stress in *Arabidopsis thaliana*. *The Plant Journal*, *77*(4), 604–615.
- Cappello, S., Denaro, R., Genovese, M., Giuliano, L., & Yakimov, M. M. (2007). Predominant growth of *Alcanivorax* during experiments on “oil spill bioremediation” in mesocosms. *Microbiological research*, *162*(2), 185–190.
- Carr, D. B., & Pickle, L. W. (2010). *Visualizing data patterns with micromaps*. CRC Press.

- Caspi, R., Altman, T., Billington, R., Dreher, K., Foerster, H., Fulcher, C. A., & others. (2014). The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases. *Nucleic acids research*, 42(D1), D459–D471.
- Chabot, C. (2009). Demystifying visual analytics. *IEEE computer graphics and applications*, 29(2), 84–87.
- Chang, R., Ziemkiewicz, C., Green, T. M., & Ribarsky, W. (2009). Defining insight for visual analytics. *Computer Graphics and Applications, IEEE*, 29(2), 14–17.
- Chen, W., Honma, K., Sharma, A., & Kuramitsu, H. K. (2006). A universal stress protein of *Porphyromonas gingivalis* is involved in stress responses and biofilm formation. *FEMS microbiology letters*, 264(1), 15–21.
- Claessen, D., Rozen, D. E., Kuipers, O. P., Søggaard-Andersen, L., & van Wezel, G. P. (2014). Bacterial solutions to multicellularity: a tale of biofilms, filaments and fruiting bodies. *Nature Reviews Microbiology*, 12(2), 115–124.
- Consortium, G. O. (2004). The Gene Ontology (GO) database and informatics resource, 32, 258–261.
- COTTINGHAM, B. (2012). Bioinformatic Strategies: Integrating Data and Knowledge to Improve Biosurveillance [Online]. Available: <http://www.dtic.mil/ndia/2012bio/Cottingham.pdf>.
- Crack, P. J., Zhang, M., Morganti-Kossmann, M. C., Morris, A. J., Wojciak, J. M., Fleming, J. K., & others. (2014). Anti-lysophosphatidic acid antibodies improve traumatic brain injury outcomes. *J Neuroinflammation*, 11, 37.
- Dasgupta, D., Ghosh, R., & Sengupta, T. K. (2013). Biofilm-mediated enhanced crude oil degradation by newly isolated *Pseudomonas species*. *ISRN biotechnology*, 2013.
- Dos Santos, R. G., Martins, A. S., Torezani, E., Baptistotte, C., Farias, J. da N., Horta, P. A., Work T. M., Balazs, G. H. (2010). Relationship between fibropapillomatosis and environmental quality: a case study with *Chelonia mydas* off Brazil. *Diseases of aquatic organisms*, 89(3), 87.
- Ericson, G., Lindesjö, E., & Balk, L. (1998). DNA adducts and histopathological lesions in perch (*Perca fluviatilis*) and northern pike (*Esox lucius*) along a polycyclic aromatic hydrocarbon gradient on the Swedish coastline of the Baltic Sea. *Canadian Journal of Fisheries and Aquatic Sciences*, 55(4), 815–824.
- Fernández-Martínez, J., Pujalte, M. J., García-Martínez, J., Mata, M., Garay, E., & Rodríguez-Valera, F. (2003). Description of *Alcanivorax venustensis* sp. nov. and reclassification of *Fundibacter jadensis* DSM 12178T (Bruns and Berthe-Corti 1999) as *Alcanivorax jadensis*

- comb. nov., members of the emended genus *Alcanivorax*. *International journal of systematic and evolutionary microbiology*, 53(1), 331–338.
- Finn, R. D., Bateman, A., Clements, J., Coghill, P., Eberhardt, Y., Eddy, S. R., & others (2013). Pfam : the protein families database, *Nucl. Acids Res.*, 42 (D1): D222-D230.
- Gallardo, K., Courty, P.-E., Le Signor, C., Wipf, D., & Vernoud, V. (2014). Sulfate transporters in the plant's response to drought and salinity: regulation and possible functions. *Frontiers in plant science*, 5, 580.
- Gertler, C., Gerdt, G., Timmis, K. N., & Golyshin, P. N. (2009). Microbial consortia in mesocosm bioremediation trial using oil sorbents, slow-release fertilizer and bioaugmentation. *FEMS microbiology ecology*, 69(2), 288–300.
- Gury, J., Seraut, H., Tran, N. P., Barthelmebs, L., Weidmann, S., Gervais, P., & Cavin, J.-F. (2009). Inactivation of PadR, the repressor of the phenolic acid stress response, by molecular interaction with Usp1, a universal stress protein from *Lactobacillus plantarum*, in *Escherichia coli*. *Applied and environmental microbiology*, 75(16), 5273–5283.
- Hanrahan, P., Stolte, C. & Mackinlay, J. (2009). Selecting a visual analytics application [Online]. Tableau Software. Available: <http://mkt.tableausoftware.com/files/selecting-visual-analytics-app.pdf>.
- Hao W., Zhang Z., Songnian H., & Jun Y. (2012). On the molecular mechanism of GC content variation among eubacterial genomes. *Biology Direct*, 7(2), 1-16.
- Hara, A., Baik, S., Syutsubo, K., Misawa, N., Smits, T. H. M., Van Beilen, J. B., & Harayama, S. (2004). Cloning and functional analysis of alkB genes in *Alcanivorax borkumensis* SK2. *Environmental microbiology*, 6(3), 191–197.
- Hara, A., Syutsubo, K., & Harayama, S. (2003). *Alcanivorax* which prevails in oil-contaminated seawater exhibits broad substrate specificity for alkane degradation. *Environmental Microbiology*, 5(9), 746–753.
- Harayama, S., Kasai, Y., & Hara, A. (2004). Microbial communities in oil-contaminated seawater. *Current opinion in biotechnology*, 15(3), 205–214.
- Heermann, R., Lippert, M., & Jung, K. (2009). Domain swapping reveals that the N-terminal domain of the sensor kinase KdpD in *Escherichia coli* is important for signaling, *BMC Microbiology*, 9(133), 1–12.
- Hildebrand, F., Meyer, A., & Eyre-walker, A. (2010). Evidence of Selection upon Genomic GC-Content in Bacteria. *PLoS GENETICS*, 6(9), 1–9.

- Hingley-Wilson, S. M., Loughheed, K. E. A., Ferguson, K., Leiva, S., & Williams, H. D. (2010). Individual *Mycobacterium tuberculosis* universal stress protein homologues are dispensable in vitro. *Tuberculosis*, *90*(4), 236–244.
- Huynen, M., Snel, B., Iii, W. L., & Bork, P. (2000). Predicting Protein Function by Genomic Context : Quantitative Evaluation and Qualitative Inferences, *Genome Res.* *10*(8):1204-10.
- Isokpehi, R. D., Mahmud, O., Mbah, A. N., Simmons, S. S., Avelar, L., Rajnarayanan, R. V, ... others. (2011). Developmental regulation of genes encoding universal stress proteins in *Schistosoma mansoni*. *Gene regulation and systems biology*, *5*, 61.
- Isokpehi, R. D., Simmons, S. S., Cohly, H. H. P., Ekunwe, S. I. N., Begonia, G. B., & Ayensu, W. K. (2011b). Identification of drought-responsive universal stress proteins in viridiplantae. *Bioinformatics and biology insights*, *5*, 41.
- Johnson, M. O., Cohly, H. H. P., Isokpehi, R. D., & Awofolu, O. R. (2010). The case for visual analytics of arsenic concentrations in foods. *International journal of environmental research and public health*, *7*(5), 1970–1983.
- Jung, H. (2002). The sodium/substrate symporter family: structural and functional features. *FEBS letters*, *529*(1), 73–77.
- Kalscheuer, R., Stöveken, T., Malkus, U., Reichelt, R., Golyshin, P. N., Sabirova, J. S., Ferrer M., Timmis K. N., Steinbüchel, A. (2007). Analysis of storage lipid accumulation in *Alcanivorax borkumensis*: evidence for alternative triacylglycerol biosynthesis routes in bacteria. *Journal of bacteriology*, *189*(3), 918–928.
- Kasai, Y., Kishira, H., Sasaki, T., Syutsubo, K., Watanabe, K., & Harayama, S. (2002). Predominant growth of *Alcanivorax* strains in oil- contaminated and nutrient-supplemented sea water, *Environ Microbiol.* *4*(3):141-7..
- Keim, D. A., Zhang, L., Krstajic, M., & Simon, S. (2012). Solving Problems with Visual Analytics: Challenges and Applications. In *ECML/PKDD (1)* (pp. 5–6).
- Keim, D., Andrienko, G., Fekete, J., Carsten, G., Kohlhammer J., Melançon G., (2008). Visual Analytics : Definition , Process , and Challenges in *Information Visualization*, *4950*: 154-175.
- Kube, M., Chernikova, T. N., Al-Ramahi, Y., Beloqui, A., Lopez-Cortez, N., Guazzaroni, M.-E., & others. (2013). Genome sequence and functional genomic analysis of the oil-degrading bacterium *Oleispira antarctica*. *Nature communications*, *4*: 2156 DOI: [10.1038/ncomms3156](https://doi.org/10.1038/ncomms3156).
- Kvint, K., Nachin, L., Diez, A., & Nyström, T. (2003). The bacterial universal stress protein: function and regulation. *Current opinion in microbiology*, *6*(2), 140–145.

- Kyrpides, N. C., Hugenholtz, P., Eisen, J. A., Woyke, T., Göker, M., Parker, C. T., & others. (2014). Genomic encyclopedia of bacteria and archaea: sequencing a myriad of type strains. *PLoS Biol* 12(8): e1001920. doi: 10.1371/journal.pbio.1001920
- Lai, Q., Wang, L., Liu, Y., Fu, Y., Zhong, H., Wang, B., Sun, F. (2011a). *Alcanivorax pacificus* sp. nov., isolated from a deep-sea pyrene-degrading consortium, *Int J Syst Evol Microbiol.* 61(6): 1370-4.
- Lai, Q., Wang, J., Gu, L., & Zheng, T. (2013b). *Alcanivorax marinus* sp. nov., isolated from deep-sea water, *Int J Syst Evol Microbiol.* 63(12): 4428-32.
- Lim C., Lee J., Choi C., Kilman V. L., Kim J., Park S. M., Jang S. K., Allada R. & Choe J. (2011). The novel gene twenty-four defines a critical translational step in the *Drosophila* clock. *Nature*, 470 (7334), 399–403.
- Liu, C., & Shao, Z. (2005). *Alcanivorax dieselolei* sp. nov., a novel alkane-degrading bacterium isolated from sea water and deep-sea sediment, *Int J Syst Evol Microbiol.* 55(3):1181-6.
- Liu, G. E. (2009). Applications and case studies of the next-generation sequencing technologies in food, nutrition and agriculture. *Recent patents on food, nutrition & agriculture*, 1(1), 75–79.
- Luan, X., Cui, Z., Gao, W., Li, Q., Yin, X., & Zheng, L. (2014). Genome sequence of the petroleum hydrocarbon-degrading bacterium *Alcanivorax* sp. strain 97CO-5. *Genome announcements*, 2(6), e01277–14.
- Mangalappalli-Illathu, A. K., & Korber, D. R. (2006). Adaptive resistance and differential protein expression of *Salmonella enterica* serovar Enteritidis biofilms exposed to benzalkonium chloride. *Antimicrobial agents and chemotherapy*, 50(11), 3588–3596.
- Marchler-Bauer, A., Lu, S., Anderson, J. B., Chitsaz, F., Derbyshire, M. K., DeWeese-Scott, C., & others. (2011). CDD: a Conserved Domain Database for the functional annotation of proteins. *Nucleic acids research*, 39(1), D225–D229.
- Markowitz, V. M., Chen, I. A., Chu, K., Szeto, E., Palaniappan, K., Pillay, M., & others. (2014). IMG / M 4 version of the integrated metagenome comparative analysis system, *Nucleic Acids Res.* 42: D568-73.
- Markowitz, V. M., Chen, I. A., Palaniappan, K., Chu, K., Szeto, E., Grechkin, Y., & Kyrpides, N. C. (2012). IMG : the integrated microbial genomes database and comparative analysis system, *Nucleic Acids Res.* 2012 40: D115-22
- Markowitz, V. M., Chen, I.-M. A., Chu, K., Szeto, E., Palaniappan, K., Grechkin, Y., & others. (2012). IMG/M: the integrated metagenome data management and comparative analysis system. *Nucleic acids research*, 40(D1), D123–D129.

- Mason, O. U., Scott, N. M., Gonzalez, A., Robbins-Pianka, A., Bælum, J., Kimbrel, J., & others. (2014). Metagenomics reveals sediment microbial community response to Deepwater Horizon oil spill. *The ISME journal*, 8(7), 1464–1475.
- Mavromatis, K., Chu, K., Ivanova, N., Hooper, S. D., Markowitz, V. M., & Nikos, C. (2009). Gene Context Analysis in the Integrated Microbial Genomes (IMG) Data Management System, *PLoS One* 4(11): e7979.
- Mbah, A. N., Mahmud, O., Awofolu, O. R., & Isokpehi, R. D. (2013). Inferences on the biochemical and environmental regulation of universal stress proteins from Schistosomiasis parasites. *Advances and applications in bioinformatics and chemistry: AABC*, 6, 15.
- McGenity, T. J. (2010). Halophilic hydrocarbon degraders. In *Handbook of hydrocarbon and lipid microbiology* (pp. 1939–1951). Springer.
- Miura, T., Tsuchikane, K., Numata, M., Hashimoto, M., Hosoyama, A., Ohji, S., Yamazoe, A., Fujita, N. (2014). Complete genome sequence of an alkane degrader, *Alcanivorax sp.* strain NBRC 101098. *Genome announcements*, 2(4), e00766–14.
- Nachin, L., Nannmark, U., & Nyström, T. (2005). Differential roles of the universal stress proteins of *Escherichia coli* in oxidative stress resistance, adhesion, and motility. *Journal of bacteriology*, 187(18), 6265–6272.
- Naether, D. J., Slawtschew, S., Stasik, S., Engel, M., Wick L. Y. , Timmis K. N., Heipieper H. J. (2013). Adaptation of the Hydrocarbonoclastic Bacterium *Alcanivorax borkumensis* SK2 to Alkanes and Toxic Organic Compounds : a Physiological and Transcriptomic Approach. *Appl Environ Microbiol.* 79(14): 4282-93
- Nakamura, S., Yang, C., Sakon, N., Ueda, M., Tougan, T., Goto, N., & others. (2009). Direct Metagenomic Detection of Viral Pathogens in Nasal and Fecal Specimens Using an Unbiased High- Throughput Sequencing Approach, *PLoS One.* 4(1):e4219.
- NATIONAL RESEARCH COUNCIL. (2010). Committee on Scientific Milestones for the Development of a Gene Sequence-Based Classification System for the Oversight of Select Agents. Sequence-Based Classification of Select Agents: A Brighter Line. 4, Committee Findings and Conclusions. [Online]. *Washington (DC): National Academies Press (US)*. Available: <http://www.ncbi.nlm.nih.gov/books/NBK50870/>.
- Nicolotti, G., & Egli, S. (1998). Soil contamination by crude oil: impact on the mycorrhizosphere and on the revegetation potential of forest trees. *Environmental pollution*, 99(1), 37–43.
- Nwankwo, C. A., Stentiford, E. I., & Fletcher, L. A. (2014). Use of Compost to Enhance the Growth of Tomatoes in Soil Contaminated with Nigerian Crude Oil. *Journal of Applied Sciences*, 14(19), 2391–2395.

- Okuda, S., Kawashima, S., Kobayashi, K., Ogasawara, N., Kanehisa, M., & Goto, S. (2007). Characterization of relationships between transcriptional units and operon structures in *Bacillus subtilis* and *Escherichia coli*. *Bmc Genomics*, 8(1), 48.
- Overholt, W. A., Green, S. J., Marks, K. P., Venkatraman, R., Prakash, O., & Kostka, J. E. (2013). Draft genome sequences for oil-degrading bacterial strains from beach sands impacted by the Deepwater Horizon oil spill. *Genome announcements*, 1(6), e01015–13.
- Parte, A. C., (2014). LPSN — list of prokaryotic names with standing in nomenclature, *Nucleic Acids Res.* 42: D613–D616.
- Peters, B. M., Jabra-Rizk, M. A., Scheper, M. A., Leid, J. G., Costerton, J. W., & Shirtliff, M. E. (2010). Microbial interactions and differential protein expression in *Staphylococcus aureus*-*Candida albicans* dual-species biofilms. *FEMS Immunology & Medical Microbiology*, 59(3), 493–503.
- Peterson, C. H., Kennicutt II, M. C., Green, R. H., Montagna, P., Harper Jr, D. E., Powell, E. N., & Roscigno, P. F. (1996). Ecological consequences of environmental perturbations associated with offshore hydrocarbon production: a perspective on long-term exposures in the Gulf of Mexico. *Canadian Journal of Fisheries and Aquatic Sciences*, 53(11), 2637–2654.
- Piłyk, S., & Paszewski, A. (2009). Sulfate permeases phylogenetic diversity of sulfate transport, *Acta Biochim Pol.* 56(3): 375-84.
- Pitcher, R. S., & Watmough, N. J. (2004). The bacterial cytochrome cbb3 oxidases, *Biochim Biophys Acta.* 1655(1-3): 388-99.
- Qiao N., and Shao Z., (2010) Isolation and characterization of a novel biosurfactant produced by hydrocarbon-degrading bacterium *Alcanivorax dieselolei* B-5. *Journal of Applied Microbiology* 108, 1207–1216
- Rahul, K, Sasikala, C., Tushar, L., Debadrita, R., & Ramana, C. V. (2014). *Alcanivorax xenomutans* sp . nov., a hydrocarbonoclastic bacterium isolated from a shrimp cultivation pond, *Int J Syst Evol Microbiol.* 64(10):3553-8
- Reva, O. N., Hallin, P. F., Willenbrock, H., Sicheritz-Ponten, T., Tümmler, B., & Ussery, D. W. (2008). Global features of the *Alcanivorax borkumensis* SK2 genome. *Environmental microbiology*, 10(3), 614–625.
- Rivas, R., García-Fraile, P., Peix, A., Mateos, P. F., Martínez-Molina, E., & Velázquez, E. (2007). *Alcanivorax balearicus* sp. nov., isolated from Lake Martel. *International journal of systematic and evolutionary microbiology*, 57(6), 1331–1335.

- Sabirova, J. S., Chernikova, T. N., Timmis, K. N., & Golyshin, P. N. (2008). Niche-specificity factors of a marine oil-degrading bacterium *Alcanivorax borkumensis* SK2. *FEMS FEMS Microbiol Lett.* 285(1): 89-96
- Sabirova, J. S., Ferrer, M., Regenhardt, D., Timmis, K. N., & Golyshin, P. N. (2006). Proteomic insights into metabolic adaptations in *Alcanivorax borkumensis* induced by alkane utilization. *Journal of bacteriology*, 188(11), 3763–3773.
- Martins dos. Santos V., Sabirova J., Timmis K. N., Yakimov M. M., Golyshin P. N., *Alcanivorax borkumensis* in *Handbook of Hydrocarbon and Lipid Microbiology* pp 1265-1288
- Schneiker S, Martins dos Santos, V. A., Bartels, D., Bekel, T., Brecht, M., Buhrmester, J., & others. (2006). Genome sequence of the ubiquitous hydrocarbon-degrading marine bacterium *Alcanivorax borkumensis*. *Nature biotechnology*, 24(8), 997–1004.
- Scholtz, J. (2006). Beyond usability: Evaluation aspects of visual analytic environments. In *Visual Analytics Science and Technology, 2006 IEEE Symposium On* (pp. 145–150).
- Schreiber, K., Boes, N., Eschbach, M., Jaensch, L., Wehland, J., Bjarnsholt, T., & others. (2006). Anaerobic survival of *Pseudomonas aeruginosa* by pyruvate fermentation requires an Usp-type stress protein. *Journal of bacteriology*, 188(2), 659–668.
- Schweikhard, E. S., Kuhlmann, S. I., Kunte, H.-J., Grammann, K., & Ziegler, C. M. (2010). Structure and function of the universal stress protein TeaD and its role in regulating the ectoine transporter TeaABC of *Halomonas elongata* DSM 2581T. *Biochemistry*, 49(10), 2194–2204.
- Shabb, J. B., Corbin, J. D., Thomas, P. J., Shenbagamurthi, P., Sondek, J., Hullihen, J. M., & others. (1992). Cyclic nucleotide-binding domains in proteins having diverse functions. *J Biol Chem.* 267(9): 5723-6.
- Sierra-García, I. N., Correa Alvarez, J., de Vasconcellos, S. P., de Souza, A., dos Santos Neto, E. V., & de Oliveira, V. M. (2014). New hydrocarbon degradation pathways in the microbial metagenome from Brazilian petroleum reservoirs. *PLoS ONE* 9(2): e90087.
- Simmons, S. S., Isokpehi, R. D., Brown, S. D., Mcallister, D. L., Rajnarayanan, R. V., & others. (2011). Bioinformatics and Biology Insights Functional Annotation Analytics of *Rhodospseudomonas palustris* Genomes. *Bioinform Biol Insights.* 5: 115–129.
- Sims, J. N., Isokpehi, R. D., Cooper, G. A., Bass, M. P., Brown, S. D., St John, A. L., Gulig P. A. Cohly, H. H. P. (2011). Visual analytics of surveillance data on foodborne vibriosis, United States, 1973-2010. *Environ Health Insights.* 5: 71–85
- Sousa, M. C., & McKay, D. B. (2001). Structure of the Universal Stress Protein of *Haemophilus influenzae*, *Structure.* 9(12): 1135-41.

- Spiegelhalter, F., & Bremer, E. (1998). Osmoregulation of the opuE proline transport gene from *Bacillus subtilis*: contributions of the sigma A- and sigma B-dependent stress-responsive promoters, *Mol Microbiol.* 29(1): 285-96..
- Szklarczyk, D., Franceschini, A., Kuhn, M., Simonovic, M., Roth, A., Minguéz, P., & others (2011). The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. *Nucleic Acids Res.* 39: D561-8
- Szklarczyk, D., Franceschini, A., Wyder, S., Forslund, K., Heller, D., Huerta-Cepas, J., & others. (2014). STRING v10: protein--protein interaction networks, integrated over the tree of life. *Nucleic Acids Res.* 43: D447-52.
- Thomas, J. J., Cook, K., & others. (2006). A visual analytics agenda. *Computer Graphics and Applications, IEEE,* 26(1), 10–13.
- Troup, B., Jahn, M., Hungerer, C., Jahn, D. (1994). Isolation of the hemF Operon Containing the Gene for the *Escherichia coli* Aerobic Coproporphyrinogen III Oxidase by In Vivo Complementation of a Yeast HEM13 Mutant, *J Bacteriol* 176(3), 673–680.
- Varghese, N. J., Mukherjee, S., Ivanova, N., Konstantinidis, K. T., Mavrommatis, K., Kyrpides, N. C., & Pati, A. (2015). Microbial species delineation using whole genome sequences. *Nucleic Acids Res.* 43(14): 6761-71.
- Wang, M., & Caetano-Anollés, G. (2009). The evolutionary mechanics of domain organization in proteomes and the rise of modularity in the protein world. *Structure,* 17(1), 66–78.
- Wang, W., & Shao, Z. (2013). Enzymes and genes involved in aerobic alkane degradation. *Front Microbiol.* 4:116
- Williams, B. S., Isokpehi, R. D., Mbah, A. N., Hollman, A. L., Bernard, C. O., Simmons, S. S., Ayensu W. K., Garner, B. L. (2012). Functional annotation analytics of bacillus genomes reveals stress responsive acetate utilization and sulfate uptake in the biotechnologically relevant bacillus megaterium. *Bioinform Biol Insights.* 6:275-86.
- Wu, Y., Lai, Q., Zhou, Z., Qiao, N., Liu, C., & Shao, Z. (2009). *Alcanivorax hongdengensis* sp. nov., an alkane-degrading bacterium isolated from surface seawater of the straits of Malacca and Singapore, producing a lipopeptide as its biosurfactant. *International journal of systematic and evolutionary microbiology,* 59(6), 1474–1479.
- Yakimov, M. M., Denaro, R., Genovese, M., Cappello, S., D’Auria, G., Chernikova, T. N., Timmis, K. N., Golyshin, P. N., Giluliano, L. (2005). Natural microbial diversity in superficial sediments of Milazzo Harbor (Sicily) and community successions during microcosm enrichment with various hydrocarbons. *Environmental microbiology,* 7(9), 1426–1441.

- Yakimov, M. M., Golyshin, P. N., Lang, S., Moore, E. R. B., Abraham, W., Lunsdorf, H., & Timmis, K. N. (1998). *Alcanivorax borkumensis* gen. nov., sp. nov., a new , hydrocarbon-degrading and surfactant-producing marine bacterium, *48*, 339–348.
- Yakimov, M. M., Timmis, K. N., & Golyshin, P. N. (2007). Obligate oil-degrading marine bacteria. *Current opinion in biotechnology*, *18*(3), 257–266.
- Zarembinski, T. I., Hung, L.W., Mueller-Dieckmann, H.J., Kim, K.K., Yokota, H., Kim, R., & Kim, S.H. (1998). Structure-based assignment of the biochemical function of a hypothetical protein: a test case of structural genomics. *Proceedings of the National Academy of Sciences*, *95*(26), 15189–15193.
- Zeilstra-ryalls, J. H., Schornberg, K. L., (2006). Analysis of hemF Gene Function and Expression in *Rhodobacter sphaeroides* 2.4.1., *J Bacteriol.* *188* (2):801-4.
- Zhang, X.Z., Xie, J.J., & Sun, F.L. (2014). Effects of three polycyclic aromatic hydrocarbons on sediment bacterial community. *Current microbiology*, *68*(6), 756–762.