

American University in Cairo

## AUC Knowledge Fountain

---

Theses and Dissertations

Student Research

---

Winter 1-31-2021

### Novel DNA ligases from the Red Sea brine pools: Cloning, expression, in silico characterization and comparative thermostability

Iyanu Oduwole  
oduwoleianu@aucegypt.edu

Follow this and additional works at: <https://fount.aucegypt.edu/etds>



Part of the [Biotechnology Commons](#)

---

#### Recommended Citation

##### APA Citation

Oduwole, I. (2021). *Novel DNA ligases from the Red Sea brine pools: Cloning, expression, in silico characterization and comparative thermostability* [Master's Thesis, the American University in Cairo]. AUC Knowledge Fountain.

<https://fount.aucegypt.edu/etds/1614>

##### MLA Citation

Oduwole, Iyanu. *Novel DNA ligases from the Red Sea brine pools: Cloning, expression, in silico characterization and comparative thermostability*. 2021. American University in Cairo, Master's Thesis. *AUC Knowledge Fountain*.

<https://fount.aucegypt.edu/etds/1614>

This Master's Thesis is brought to you for free and open access by the Student Research at AUC Knowledge Fountain. It has been accepted for inclusion in Theses and Dissertations by an authorized administrator of AUC Knowledge Fountain. For more information, please contact [thesisadmin@aucegypt.edu](mailto:thesisadmin@aucegypt.edu).



School of Sciences and Engineering

**Novel DNA ligases from the Red Sea brine pools: Cloning, expression, in silico characterization and comparative thermostability**

A thesis submitted to  
Biotechnology graduate's program  
In partial fulfillment of the requirements  
For the degree of Master of Science

By

Oduwole Iyanu Mumeen

Under the Supervision of:

Prof. Rania Siam

Professor, Biology Department

Co supervisor:

Dr. Rehab Abdallah

Adjunct Assistant Professor and Post-doctoral Researcher, Biology Department

September 2020

**Novel DNA ligases from the Red Sea brine pools: Cloning, expression, in silico  
characterization and comparative thermostability**

A Thesis submitted by

Oduwole Iyanu Mumeen

To Biotechnology Graduate program

In partial fulfillment of the requirements for

The Master of Science degree

Has been approved by

Thesis Committee Supervisor/Chair \_\_\_\_\_

Affiliation \_\_\_\_\_

Thesis Co-advisor \_\_\_\_\_

Affiliation \_\_\_\_\_

Thesis Committee Reader/Examiner \_\_\_\_\_

Affiliation \_\_\_\_\_

Thesis Committee Reader/Examiner \_\_\_\_\_

Affiliation \_\_\_\_\_

\_\_\_\_\_  
**Dept. Chair/Director**

\_\_\_\_\_  
**Date**

\_\_\_\_\_  
**Dean**

## **Dedication**

I am dedicating this thesis to my late father, Mr. Rasheed Tomola Oduwole who was a highly disciplined teacher, and to my lovable mum whose support in my life is endless. I am also extending this dedication to my siblings ( Lola, Dare, Kafilat, Seun and Sakiru) and many friends at home, in AUC and the community who have always shown me love, care and encouragement throughout the course of doing this thesis.

## Acknowledgement

Firs to all, I thank Allah for his infinite mercy, protection and guidance in my life. This thesis would never be completed without the help of my advisor Prof. Rania Siam and my co-advisor. Dr. Rehab Abdallah. Prof. Rania has been not only my advisor but also my mentor, and she has been there with me through easy and challenging times at AUC. I do not think I would ever forget the contributions of Dr. Rehab in my thesis and research. She is such a hardworking and brilliant person who has always stimulated me to learn and grow, and I am confident she is going to be a fantastic professor in the future. I also appreciate Dr. Ahmed Moustafa who adequately exposed me to bioinformatics and computational works that assisted greatly in doing the bioinformatics part of my project. I would like to thank the office of Associate Provost for Research Innovation and Creativity for granting the AUC centennial grant “Proof Concept fund” to Dr. Rania which supported my research together with AUC graduate research grant.

My deepest gratitude also goes to Prof. Rania's research extremozymes research group. I would not ever forget the contributions of my laboratory colleague, Shaimaa Farag, genome specialist, Amged Ouf and a doctoral student, Mohammed Malash towards my research and thesis. I would forever appreciate all the AUC biotechnology professors (Dr. Asma Amleh, Dr. Walid Fouad, Dr. Anwar Abdel Elnasser, Dr. Ahmed Moustafa and Dr. Ahmed Abdellatif) who have taught me during my master's program. I am also extending my appreciation to the lab aides for their consistent assistance during my bench work. Besides, I would not be in AUC without winning the African graduate fellowship which is why I would forever appreciate the Biotechnology admission committee and university fellowships who deem me fit for admission and fellowship. However, I would specially thank the dean of the graduate studies, Dr. Adham Ramadan, the sub-dean Mrs. Aya Morsi, Mohammed Wael, Salma Serry, and Nashwa for their supports throughout my stay in AUC.

I will forever remember the love and support of all the biotechnology graduates' students, the African Heritage Association (AHA) members, Graduate Student Association (GSA) members and my friends. More so, I would not fail to appreciate some friends; Khalid, Fahad, Sis Aisha, Mum Sis Aisha, Abdelhameed, Mahmoud, Kathir, Samuel, Paul, Uzor, Chimson, Gehad, Amira, Logayn, Mie and Neela and many more. They have all made my sojourn in AUC and Egypt a wholesome one. I also appreciate all the AUC staff, particularly the residential receptionist Mr. Ehssan for his support and kindness always. Lastly, I will forever remember the love and kindness shown to me by Sheikh Dr. Habeebulah Hashim Al imam and his family throughout my stay in Egypt.

## Abstract

The American University in Cairo

### **Novel DNA ligases from the Red Sea brine pools: Cloning, expression, in silico characterization and comparative thermostability.**

Oduwole Iyanu Mumeen

Supervisors: Prof. Rania Siam, Dr. Ahmed Moustafa, and Dr. Rehab Abdallah

---

Extreme physicochemical conditions such as high temperature, salinity, and the presence of heavy metal are characteristics of some of the Red Sea brine pools environment. We screened two Red Sea Brine pools (Atlantis II(AT-II), and Discovery Deeps (DD), and one interface layer (Kebrit Deep) to identify novel DNA ligases with potential extreme biochemical properties. Furthermore, we did an in silico comparative thermostability study by examining the stability role of proline and arginine residues at the loop conformations and exposed regions of ligase sequences from metagenomic assemblies of different extreme environments, including the Red Sea metagenomes.

A sequence-based metagenomics approach was used to identify the putative DNA ligase sequences from the Red Sea brine pools and interface layer metagenomes downloaded from the NCBI database. 6, 148, 453 metagenomic reads were assembled using MEGAHIT, which generated 783,176 contigs. A concatenated HMM model built from raw HMM models of ATP and NAD<sup>+</sup> ligases domains available from the Pfam database was used to scan predicted ORFs from contigs. A total of 18 ORFs were identified, and two of the ORFs, LigATL1 (ATP type), from AT-II and LigKDU4 (NAD<sup>+</sup> type) from KB, were selected for synthesis, phylogenetic study, and further preliminary characterizations. LigATL1 was cloned, expressed, and partially purified. Additionally, ligase sequences from psychrophilic, mesophilic, thermophilic, and hyperthermophilic environments were retrieved from the NCBI database for comparative thermostability study with some of the putative Red Sea ligase sequences. The retrieved 22 ligase sequences were divided into five different closest taxonomic groups. ConSurf and DisEMBL servers were used to analyze Proline (Pro) and Arginine (Arg) residues in the exposed/buried regions and the loop and hot loops regions of the putative ligases (retrieved + Red Sea), respectively.

A putative LigATL1 showed a 38% identity to ATP-Dependent DNA ligase from *Erysipelotrichaceae bacterium*, while LigKDU4 has a 60% identity to NAD<sup>+</sup> Dependent DNA ligase from *Candidatus Marinimicrobia bacterium*. The phylogenetic analysis suggests that LigATL1 belongs to the LigD(ATP type) family, while LigKDU4 is amongst the LigA family,(NAD<sup>+</sup> type). LigATL1 has 100% confidence modeling using bound-adenylated nicked human DNA ligase as a template, and is superimposed with the highest similarity (Template modeling <sup>TM</sup> score =1.0) to thermostable DNA ligase from *S.solfataricus*. LigKDU4 modeled with 100% confidence using bound-adenylated nicked *E.coli* DNA ligase, and also superimposed with the highest similarity(TM score= 1.0) to thermostable *t2 filiform* DNA ligase. *In vitro*, functional assay and biochemical characterization are still required to confirm both enzyme activity and thermostability.

For the comparative thermostability analysis, many Ligase sequences from thermophilic or hyperthermophilic environments had higher Pro and Arg residues both at the exposed and the hot loops regions than those from other mesophilic and psychrophilic environments. The highest buried Pro and Arg residues were reported for ligase sequences from psychrophilic environments at almost all the groups. Two out of five putative ligase sequences selected for the thermophilic AT-II environment had more hot loops and less buried Pro and Arg residues than other pairs in their respective groups. In the case of LigKDU4(MLK), it has the highest hot loop and exposed Arg residues than its pairs in its group which is unusual when compared to Arg analysis in other groups. This comparative study can give an insight into improving the thermal stability of enzymes generally.

## TABLE OF CONTENTS

List of Figures .....	ix
List of Tables .....	xii
List of Abbreviations .....	xiv
Chapter 1: Literature Review .....	1
1. Red Sea Brine pools: Origin, Evolution, and Microbiology .....	1
1.1 Atlantis II Deep brine pools .....	1
1.2 Kebrit Deep brine pools .....	2
1.3 Other Red Sea Brine pools .....	3
2. Metagenomics Approach For Mining unique biocatalysts .....	4
2.1 Construction of Metagenomic Libraries .....	4
2.2 Screening of Constructed Libraries for Gene of Interest .....	5
2.2.1 Sequence-based Screening .....	5
2.2.2 Function-based Screening .....	5
3. DNA Ligase .....	7
3.1 Biochemical structure and functions of ATP Dependent DNA Ligases (ADLs) .....	7
3.2 Biochemical structure and functions of NAD <sup>+</sup> Dependent DNA Ligases (NDLs) .....	8
3.3 Conserved Sequences in DNA ligase .....	10
3.4 Evolutionary Relationships of Bacterial DNA Ligases .....	11
3.5 Reaction Mechanism of DNA ligase .....	12
3.6 Nick Determination and Discrimination in DNA ligases .....	13
3.7 DNA Ligase Assays .....	14
3.8 Industrial and Biotechnological Applications of DNA ligase .....	15
3.8.1 DNA ligase applications in gene cloning .....	15
3.8.2 DNA ligase as a molecular engineering tool in DNA assembly .....	15
3.8.3 Use of ligase in Next- Generation Sequencing (NGS) technologies. ....	16
3.8.4 DNA ligase' uses in DNA Origami .....	16
3.8.5 DNA ligase in Molecular Diagnostics .....	17
3.8.6 NAD <sup>+</sup> Dependent DNA ligases as chemotherapeutic targets in determining novel antibiotics candidates. ....	18
4. Rigidity, flexibility and enzyme thermostability .....	19

5. Role of Proline and Arginine in conferring thermostability to proteins.....	20
5.1 Relationship between proline and arginine residues in the disordered regions to thermostability .....	21
6. Rationale and goal of the study .....	23
Chapter 2: Materials and Methods.....	24
1. Sample Collection.....	24
2. DNA Extraction, Sequencing, and generation of Metagenomic Libraries.....	24
3. Bioinformatics Analysis I.....	24
3.1 Contig Assembly.....	24
3.2 DNA Ligase Screening from the Red Sea brine pools Metagenomic datasets.....	25
3.3 Multiple Sequence Alignment (MSA) of all the Selected DNA Ligase ORFs from Red Sea Metagenomic Datasets .....	25
3.4 Phylogenetic Analysis of Red Sea Metagenome Putative Lig ATL1 (ATP type) and LigKDU4 (NAD <sup>+</sup> type).....	26
3.5 3D Modelling and Predicting the Superimposition of LigATL1 (ATP type) and LigKDU4 (NAD <sup>+</sup> type).....	26
3.6 Prediction of Molecular weight and Isoelectric point(pI) of LigATL1 and ligKDU4 .....	26
4. Bioinformatics Analysis II: Comparative thermostability analysis of Red Sea Ligase sequences to ligase sequences of different environments; stability roles of Proline and Arginine .....	26
4.1 Retrieving different ligase sequences from the NCBI.....	26
4.2 Calculation of the Arginine and Proline compositions in the ligase sequences' primary structures .....	27
4.3 Analyzing the ligase sequences for globularity and disorder .....	27
4.4 Analyzing the buried, exposed and functional residues.....	28
4.5 Proline and Arginine residues analysis in the loops and exposed regions of ligase sequences from different environment.....	28
5. Gene Synthesis, Cloning and Transformation.....	29
6. Expression of Lig-ATL1 in <i>E coli</i> BL21(DE3) and <i>E coli</i> BL21 PLYS.....	29
7. Partial Purification of His-tagged Lig-ATL1 on Ni <sup>2+</sup> column .....	30
Chapter 3: Results and Discussions .....	31
1. Contig Assembly .....	31
2. DNA Ligase Screening from the Red Sea brine pools Metagenomic datasets .....	31
2.1 Multiple Sequence Alignment (MSA) of all the Selected DNA Ligase ORFs from Red Sea Metagenomic Datasets .....	34



2.2 Phylogenetic Analysis of LigATL1 and LigKDU4 .....	38
3. In silico Characterizations of LigATL1 and LigKDU4.....	41
3.1 3D Modelling and Predicting the Superimposition of LigATL1 (ATP type) and LigKDU4 (NAD <sup>+</sup> type).....	41
4. Predicting the Molecular weight and Isoelectric point (pI) of LigATL1 and ligKDU4.....	44
5. Comparative thermostability analysis of Red Sea ligase sequences to ligase sequences different environments: stability roles of proline and Arginine.....	44
5.1 Arginine and Proline Compositions in the Primary Structures of the ligase sequences .....	45
5.2 Analysis of Ligase segments for disorder and globularity.....	47
5.3 Analysis of proline and arginine residues in the loops, hot loops, exposed and buried regions of the ligases .....	50
6. Gene Synthesis, Cloning and Transformation of LigATL1.....	54
6.1 Cloning and Transforming of LigATLI to cloning host <i>E. coli</i> DH5 $\alpha$ .....	54
6.2 Cloning and Transforming of LigATLI in Expression hosts.....	55
6.3 Expression of Lig-ATL1.....	56
6.4 Partial Purification of His-tagged Lig-ATL1 on Ni <sup>2+</sup> column.....	60
Chapter 4: Conclusion and Recommendation.....	63
References.....	64

## List of Figures

- Figure (1): Bathymetric chart showing the locations of various red sed brine pools,..... 3**
- Figure (2): Steps involve in Next generation Sequencing Metagenomics approach,..... 6**
- Figure (3): Conserved Pfam Domains within the NAD<sup>+</sup> and ATP Dependent DNA ligases.10**
- Figure (4): Conserved Sequence elements alignment in Bacterial DNA Ligases, Six motifs have been identified in DNA ligases.. ..... 11**
- Figure (5): Mechanism of Reaction for NAD<sup>+</sup> Dependent DNA ligases..... 13**
- Figure (6): Propensities of amino acids to be disordered according to Russell/ Linding hot loops and loops definitions. .... 22**
- Figure (7): The Pipeline for screening DNA ligases from the assembled Red Sea Metagenomic datasets..... 25**
- Figure (8): Essential Motifs in ORFs coding for Putative ATP Dependent DNA Ligases from the Red Sea Metagenomes:..... 35**
- Figure (9): Multiple Sequence Alignment showing the essential Motifs in ORFs coding for Putative NAD<sup>+</sup> Dependent DNA Ligases from the Red Sea Metagenomes: ..... 36**
- Figure (10): Alignment of LigKDU4 and LigATL1 with the conserved sequences among Bacterial DNA ligases ( NAD<sup>+</sup> type to the right and ATP type to the left)...... 37**
- Figure (11a): Phylogenetic analysis for LigATL1 and LigKDU4..... 39**
- Figure 11b: Reconstructed phylogenetic tree for LigATL sequence. .... 40**
- Figure (12): Modelling of LigATL1 using a crystal structure of human DNA ligase bound to 5'-adenylated, nicked2 DNA as a template..... 41**

<b>Figure (13): Modelling of ORFKDU4 with NAD+ <i>E.coli</i> DNA ligase bound2 to nicked DNA-adenylate .....</b>	<b>42</b>
<b>Figure (14): Superimposition of LigATL1 with thermostable ATP-dependent DNA ligase from <i>S. solfataricus</i>,.....</b>	<b>43</b>
<b>Figure (15): Superimposition of LigKDU4 (Kebrit Deep) with thermostable <i>t2.filliformis</i> DNA Ligase,.....</b>	<b>43</b>
<b>Figure (16): Summary of Proline residue analysis at the hot loops, exposed functional and buried regions.....</b>	<b>52</b>
<b>Figure 17: Summary of Arginine residue analysis at the hot loops, Exposed functional and buried regions.....</b>	<b>53</b>
<b>Figure (18) : Restriction digestion of recombinant pUC19 containing ligase gene insert at NdeI and Bam HI sites. Lane 1; ; GeneRuler™ 100bp plus DNA ladder (Thermoscientific), lane 2; pUC 19 digested with NdeI and Bam HI.....</b>	<b>54</b>
<b>Figure (19): Restriction digestion of recombinant pET 16B containing ligase gene insert..</b>	<b>55</b>
<b>Figure (20): Restriction digestion of recombinant pET 16b containing ligase gene insert to confirm orientation of the insert. ....</b>	<b>56</b>
<b>Figure (21) : SDS-PAGE Analysis of LigATL1 following the expression in <i>E.coli</i> BL21 DE3 induction condition of 6hrs at 37 °C, lysis buffer pH 7.4, and IPTG conc. 0.4mM and 0.5mM NaCl.....</b>	<b>57</b>
<b>Figure (22): SDS-PAGE Analysis of LigATL1 following the expression in <i>E.coli</i> BL21 DE3 and <i>E.coli</i> PLys induction condition of 3hrs at 37 °C and overnight (16hrs at 16°C), lysis buffer pH 7.4, IPTG conc. 0.4mM and 0.5mM NaCl .....</b>	<b>58</b>
<b>Figure(23): SDS-PAGE Analysis of LigATL1 following the expression in <i>E.coli</i> BL21 DE3, induction condition of 6hrs, lysis buffer pH 8.5 &amp;6.4, IPTG conc. 0.4mM and 0.5mM NaCl .....</b>	<b>59</b>

**Figure(24) : SDS-PAGE Analysis of LigATL1 following the expression in E.coli BL21 DE3, induction condition of 5hrs using different lysis buffers( Na-Glycine buffer pH 10.0, Na-Phosphate buffer pH 8.0), different IPTG conc( 0.1& 0.4mM) and 0.3mM NaCl ..... 59**

**Figure(25): SDS-PAGE Analysis of LigATL1 following the expression in *E.coli BL21* DE3, induction condition of 5hrs using Na-Phosphate lysis buffer pH 8.0, IPTG 0.1mM and 0.3mM NaCl at large scale expression. .... 60**

**Figure (26): SDS-PAGE analysis of partial purification of LigATLI on Ni-NTA column under the native condition with binding buffer (20mM of NaH<sub>2</sub> PO<sub>4</sub>, pH 8.0. 0.3M NaCl, 10% glycerol, 0.2% Triton X and 0mM imidazole) and imidazole stepwise elution (10-500mM) ..... 61**

**Figure(27): SDS-PAGE analysis of partial purification of LigATL on Ni-NTA column under native condition with binding buffer (20mM of NaH<sub>2</sub> PO<sub>4</sub>, pH 8.0, 0.3M NaCl, 10% glycerol, 0.2% Triton X and 0mM imidazole) and stepwise elution from 30-40mM. .... 62**

## Supplementary Figures

**Fig 2. Proline and Arginine residues analysis in the primary structures, exposed and buried regions of ligase sequences belonging to the *Candidatus marinimicrobia bacterium*..... 87**

**Fig 3 Proline and Arginine residues analysis in the primary structures, loop/coils and hot loops regions of ligase sequences belonging to the *Acidimicrobiaceae bacterium* species... 88**

**Fig.4 Proline and Arginine residues analysis in the primary structures, exposed and buried regions of ligase sequences belonging to the *Acidimicrobiaceae bacterium*..... 88**

**Fig.5 Proline and Arginine residues analysis in the primary structures, loop/coils and hot loops regions of ligase sequences belonging to the *Moraxallaceae bacterium* ..... 89**

**Fig.6 Proline and Arginine residues analysis in the primary structures, loop/coils and hot loops regions of ligase sequences belonging to the *Moraxallaceae bacterium* ..... 89**

**Fig.7 Proline and Arginine residues analysis in the primary structures, loop/coils and hot loops regions of ligase sequences belonging to the *Phyllobacterium myrsinacearum* ..... 90**

**Fig.8 Proline and Arginine residues analysis in the primary structures, loop/coils and hot loops regions of ligase sequences belonging to the *Phyllobacterium myrsinacearum* species ..... 90**

**Fig.9 Proline and Arginine residues analysis in the primary structures, loop/coils and hot loops regions of ligase sequences belonging to the *Rhizobiales bacterium* species ..... 91**

**Fig.11 Proline and Arginine residues analysis in the primary structures, loop/coils and hot loops regions of ligase sequences belonging to the *Candidatus marinimicrobia bacterium* (*Partial*)..... 92**

**Fig .12 Proline and Arginine residues analysis in the primary structures, exposed and buried regions of ligase sequences belonging to the *Candidatus marinimicrobia bacterium*(*partial*) ..... 92**

## List of Tables

<b>Table 1: Environments and abbreviations of retrieved ligase sequences used in the comparative thermostability study.....</b>	<b>27</b>
<b>Table 2: Assembling of reads of Red Sea metagenomes to contigs using MEGAHIT .....</b>	<b>31</b>
<b>Table 3: Putative ORFs for DNA ligase sequences selected from the predicted ORFs of Red Sea brine pools final Contigs.....</b>	<b>32</b>
<b>3a) Putative DNA ligases (ATP type) from Atlantis II brine pool assembled final contigs .</b>	<b>32</b>
<b>3b) Putative DNA ligases (NAD<sup>+</sup> type) from Atlantis II brine pool assembled final contigs</b>	<b>33</b>
<b>3c) Putative DNA ligases from Discovery Deep brine pool (DDP) and Kebrit Deep Upper (KDU) and Kebrit Deep Lower Surfaces assembled final contigs .....</b>	<b>33</b>
<b>3d) The ORFs selected for Synthesis.....</b>	<b>34</b>
<b>Table 4: Isoelectric points and Molecular weights calculations of LigATL1 and LigKDU4 sequence .....</b>	<b>44</b>
<b>Table 5: Pro and Arg Compositions of ligase sequences used in the Comparative study.</b>	<b>46</b>
<b>Table 6: Globularity and Disordered analysis of Ligase sequences .....</b>	<b>49</b>

## List of Abbreviations

ATII Atlantis II brine pool LCL Lower convective layer  
KDU-Kebrit Deep Upper Surface  
KDL-Kebrit Deep Lower Surface  
DDP- Discovery Deep  
CD-Chain Deep  
KAUST King Abdullah University of Science and Technology  
BAC Bacterial artificial chromosome  
NGS Next-generation sequencing  
SIGX Substrate induced gene expression screening system  
NCBI National center for biotechnology information  
BLAST Basic local alignment search tool IR Ionizing radiation  
DSB Double-strand breaks  
HR Homologous recombination  
NHEJ Non -Homologous end joining  
ATP Adenosine triphosphate  
ADLs- ATP Dependent DNA ligase  
NAD Nicotinamide adenine dinucleotide  
NT Nucleotidyl transferase  
 $\beta$  NMN- Nicotinamide mononucleotide  
NDLs- NAD<sup>+</sup> Dependent DNA ligase  
OB Oligonucleotide binding  
SNP Single nucleotide polymorphism  
LCR Ligase chain reaction  
Taglig-Taq-DNA ligase  
ELISA Enzyme-linked immunosorbent assay  
Pro-Proline  
Arg-Arginine

IDP Intrinsic Disordered Protein

IDRs- Intrinsic Disordered Region

HL-Hot loops

CTD Conductivity, Temperature and Depth

ORF Open reading frame LDF Linear discriminant function

CDD Conserved domain database ML Maximum Likelihood

pI Isoelectric point

MEGA-Molecular Genetics Evolutionary Analysis

IPTG Isopropyl  $\beta$ -D-1-thiogalactopyranoside

SDS-PAGE Sodium dodecyl sulfate-polyacrylamide gel electrophoresis



## Chapter 1: Literature Review

### 1. Red Sea Brine pools: Origin, Evolution, and Microbiology

The Red Sea lies between the uplifted Arabian and African shields and has 250 to 450km coverage with elongated- bounded depression and shallow continental shelves (Cochran, 1983). The Red Sea rift system is a typical oceanic rifting process resulting from a continent break up which has successfully divided it into the southern, central, and northern regions(Cochran, 1983; Winckler et al., 2001). The north and central regions of the Red Sea were discovered in 1966 as isolated topographic depressions that contained geothermal brines, waters with high salinities, and temperatures (Cochran, 1983; Pautot et al., 1984). These brines are formed in horizontal uniform layers that have thin interfaces representing distinct gradients of temperature and salinity (Anschutz et al., 1999; Winckler et al., 2001). The Red Sea approximately has 25 deep-sea anoxic brine pools, with each having distinct, extreme, and complex ecosystems with physicochemical properties that are very hostile to the earth's environment (Antunes et al., 2011). The anoxic property of the brine pools results from a combined process of decomposed organic matter, brine reduction, limited oxygen renewal and diffusion across the interfaces of the brine seawater (Anschutz et al., 1999)

The density gradient formed from the interfaces allows the trapping of organic and inorganic materials from the seawater, thus increasing the nutrient supply, which permits microbial growth. This also created differences in physiochemical gradients between the interfaces, which also resulted in several unique microbial niches (Cronan, 1999; Eder et al., 2002). Besides, the anoxic brine can interact with the weak alkaline Red Sea waters, which facilitates the deposition of heavy metal, and this is responsible for the high content of metalliferous sediments seen in the brine pools (Ross, 1972). In recent decades, the study of the microbiology of the Red Sea has been receiving much attention, and so far, several diverse localized microbial communities have been identified. Furthermore, new taxonomic groups and novel extremophiles that are surviving in these environments have also been discovered through combined approaches of molecular and cultivation-based studies (Antunes et al., 2011). Some Red Sea brine pools include;

#### 1.1 Atlantis II Deep brine pools

Atlantis II Deep is the most dynamic, best studied, and largest Red Sea brine pool with a maximum depth of 2194m and a 200m thick filled brine having a total volume between 17-20km<sup>3</sup>. (Anschutz et al., 1999; Blanc & Anschutz, 1995). The brine is stratified into many layers, and this corresponds to a steady increase in the temperature and salinity. The Deepsea is said to be hydrothermally active due to increasing temperature within the brine with the lowest brine layer having a temperature, salinity, and pH of 68.2°C, 25.7%, and 5.3, respectively (Karbe, 1987; A. R. Miller et al., 1966). The temperature gradient splits the brine pools into four layers, which are three upper convective layers and one lower convective layer (LCL). The LCL layer is the deepest, the hottest, and the saltiest (Winckler et al., 2001). This brine pool is also connected at the

subsurface to the two adjacent brine pools: the Discovery and Chain Deeps. In addition, the brine pool has higher amounts of dissolved gas, such as Nitrogen, methane, and little amounts of carbon dioxide, hydrogen sulfide, and ethane (Backer & Schoell, 1972; Faber et al., 1998). The high content of metalliferous sediments are very rich in iron, zinc, copper and other heavy metals found in the Atlantis II deep constitute the major ore-forming body presently (Backer & Schoell, 1972; Faber et al., 1998)

In the late 1960s, the Atlantis II and Discovery Deeps were surveyed for any microbial growth. It was reported as sterile due to the absence of growth on the media tested, and the environment was suggested to be too harsh to support any form of existence (Degens & Ross, 2013; Watson & Waterbury, 1969). In contrast, sulfate-reducing microbes belonging to the halotolerant *Desulfovibrio* genus were isolated from the interface of the less-harsh brine seawater, but these organisms were not described fully (Trüper, 1969). Furthermore, *Flexistipes sinusarabici* is a Gram-negative anaerobic bacterium that was reported to have been isolated and characterized from the brine seawater interface, and this was the first microbe to be described from the Red Sea anoxic brine pools. This bacterium was later placed under a new separate phylum called Deferribacteres (Huber & Stetter, 2015; Ludwig et al., 1991). Later on, Many diverse bacterial and archaeal groups From Atlantis II brine pools have been revealed through Comparative metagenomic analysis (Abdallah et al., 2014; Antunes et al., 2011; Behzad et al., 2016; Siam et al., 2012; Wang et al., 2013).

## 1.2 Kebrit Deep brine pools

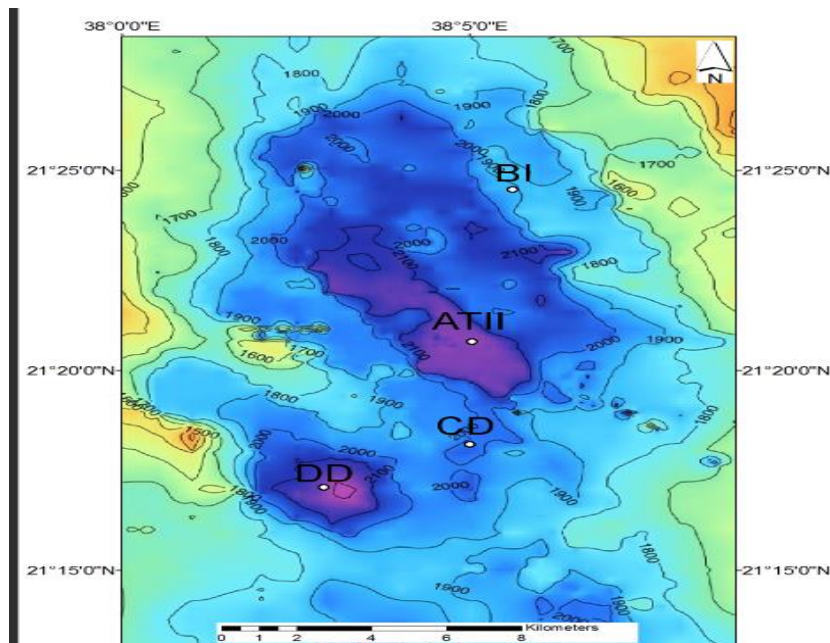
It is among the smallest brine pools of the Red Sea, has 1 by 2.5 km basin, which resembles an oval shape that is highly steeped at its sides. The Kebrit Deep has an upper boundary at 1465m water depth and a single brine layer. Despite having similarity in brine salinity to Atlantis II Deep, it differs in some of its hydrographic and geochemical characteristics (Winckler et al., 2001). The Kebrit Deep has slightly acidic brine, high salinity of 26%, and a temperature of 23.3°C (Winckler et al., 2001). The high amount of gases like Nitrogen, methane, and ethane can be found in the Kebrit Deep interfaces (Abdallah et al., 2014; Eder et al., 2001; Hartmann et al., 1998). Its high content of Hydrogen sulfide imparts the strong characteristic odor of Sulphur, and that is where the name Kebrit (the Arabic word for sulfur) was derived. These sulfides from the Kebrit Deep brine pool are derivatives of pyrite and sphalerite, which are characterized by porosity and fragility (Scholten et al., 2017).

Earlier on, microbial activities have been discovered in Kebrit Deep right before any microbiological studies due to the discovery of sulfur isotopes, which might be involved in several processes like sulfide formation, sulfate reduction, and bacterial methane oxidation. may be happening at the interface of the brine seawater (Abdallah et al., 2014; Faber et al., 1998). A novel halophilic bacterium was first identified from the phylogenetic studies of kebrit sediment sample. This sequence from this Kebrit sample was later grouped close to Aquificales and Thermotogales and was given a division number EM19 (Antunes et al., 2011; Reysenbach et al., 1994). In addition, some of the archaeal sequences from this brine pool clustered within a Thermoplasmatales group (Euryarchaeotaphylum). At the same time, the rest of the archaeal sequences showed little similarity to any phylogenetic groups (Antunes et al., 2011; Eder et al.,

2002). Many similar studies on the kebrit brine-sea water interfaces have identified sequences taxonomically assigned to Spirochetes, Clostridiales, Actinobacteria, but they did not detect any archaeal sequences. Similarly, some of the discovered sequence clusters matched to the Proteobacteria class and the iron-oxidizing bacteria *Mariprofundus ferroxidans* which are highly abundant in the Red-Sea environment (Emerson et al., 2007)

### 1.3 Other Red Sea Brine pools

Some other brine pools in the Red Sea include Discovery Deep, Shaban Deep, and Chain Deep. Discovery Deep is connected to the Atlantis II brine through the subsurface connections. However, no hydrothermal activity has been reported due to the inadequate long term documentation of temperature variations (Monnin & Ramboz, 1996). However, this Deep sea brine pool has less extreme conditions, high content of metals, moderate temperatures and is not elaborately described as Atlantis II brine pools (Antunes et al., 2011). The Shaban Deep is like Kebrit Deep but with less frequent sulfides emission, highly enriched in nickel, and absence of copper (Michaelis et al., 1990). It is 10 by 6 km in area, approximately rhombic shape, and has four different basins with a central ridge, which also has two different saddles. The upper brine of Shaban Deep has a pH and temperature of approximately 6.0 and 23°C, respectively (Hartmann et al., 1998).



**Figure (1): Bathymetric chart showing the locations of various red sed brine pools, ATII- Atlantis II brine pool, DD- Discovery Deep brine pool, CD- Chain Deep brine pool, BI- Brine Influence Sediment pool (Adapted from Siam, R. et al. 2012)**

## 2. Metagenomics Approach For Mining unique biocatalysts

Marine bioprospecting is a new area targeted towards the searching of novel organisms or genes in less explored areas such as on the deep seabed, extreme depth, cold seeps, and hydrothermal vents. Scientists always integrate High-throughput DNA sequencing with bioprospecting to evaluate marine biodiversity, and to provide a vital means to identify new enzymes and active metabolites involved in various biosynthesis (Abida et al., 2013; Arrieta et al., 2010). Many microorganisms are difficult to be cultured because they require special growth conditions. Besides, under unnatural conditions, some microorganisms will not survive because they are interdependent with other organisms (Pace, 2009). Even though cultivation based approaches have been dramatically improved, more than 70 bacteria phyla cannot still be cultured in the laboratory (Barone et al., 2014).

Omics techniques are currently used to explore genetic diversity from several sources. Metagenomics is an essential omics technique that involves mainly the direct extraction and cloning of DNA recovered from a mixture of organisms found in an environment (Handelsman, 2004). This environment could be highly diverse, artificially enriched, or extreme ones (Barone et al., 2014). New sequencing technologies such as third-generation sequencing technologies have made screening and analysis of sequences highly effective (Barone et al., 2014). Single-cell sequencing is used in sequencing individual molecules and other sequencing methods which use fluorescence detection have also been developed (Xu et al., 2009). Next-generation Sequencing Simulator for Metagenomics (NeSSM) has been useful for high-throughput metagenome sequencing (Jia et al., 2013). The enormous metagenomic data derived from sequencing process can easily be analyzed by Bio-informatics software, for instance, MEGAN (analyzing the taxonomies in large metagenomic data sets from 454 sequencing) and Meta QC (performs high QC (quality control) on the metagenomic data) (Huson et al., 2007; Q. Zhou et al., 2014).

The Metagenomics approach involves four main steps:

- A) Extraction of DNA, and construction of metagenomic libraries
- B) Screening the constructed libraries for the desired genes
- C) Functional and sequential analysis of the desired gene
- D) Expression and characterization of the gene of interest

### 2.1 Construction of Metagenomic Libraries

This step involves the insertion of large genomic DNA usually between 40 -200kb in size either through a T-A or blunt end ligation into a large cloning vector, for example, cosmids or BACs. This method ensures that large gene clusters or operons can be detected, but in a case where a gene has a low expression level, libraries may be constructed using small inserts with the plasmids. *E.coli* is commonly employed as the host strain for transformation because of ease in genetic manipulation and other downstream processes. However, *E.coli* may not adequately

express some specific genes, and in this case, other alternative cloning hosts like *Bacillus subtilis*, *Pseudomonas* spp or eukaryotic hosts may be considered (Singh et al., 2009; Uria et al., 2005)

## **2.2 Screening of Constructed Libraries for Gene of Interest**

The screening process can generally be performed through a sequence-based screening or function-based screening.

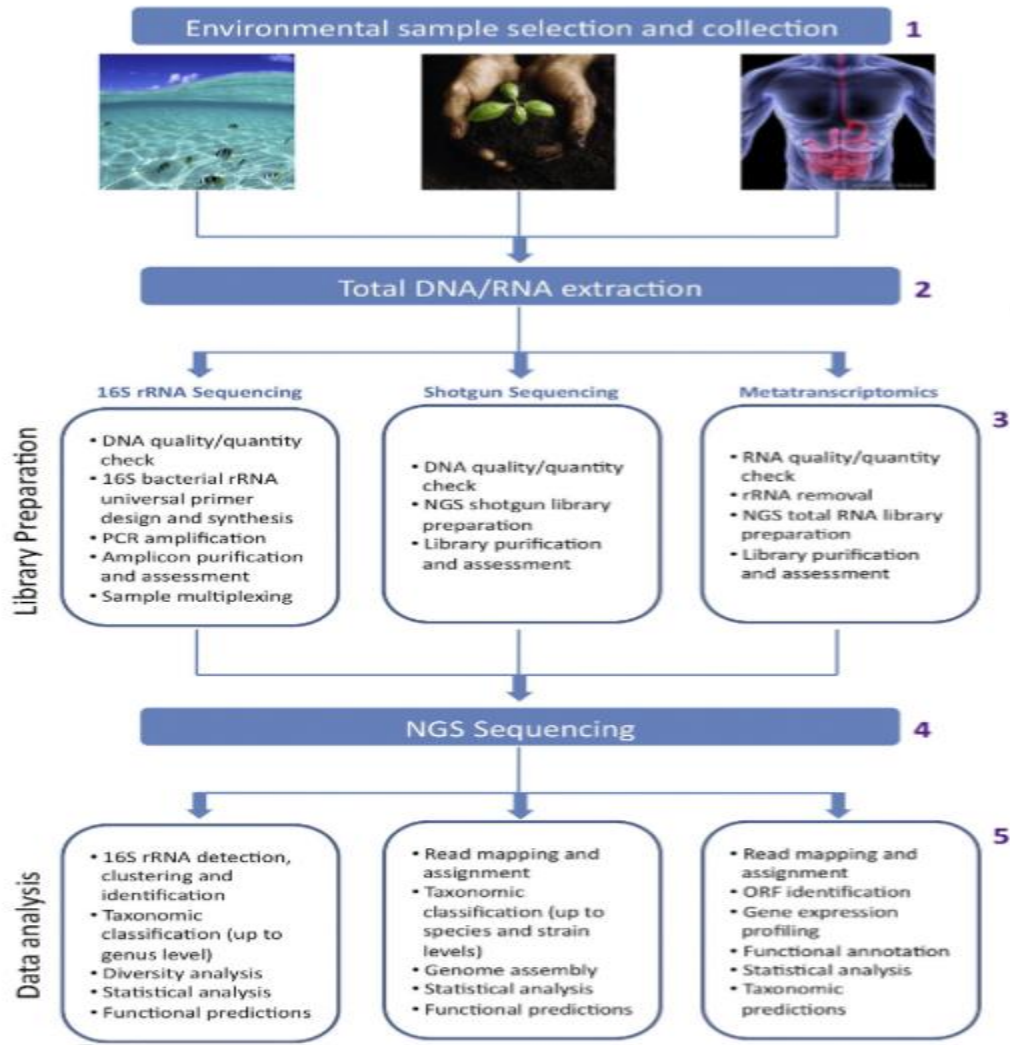
### **2.2.1 Sequence-based Screening**

This screening method was first applied to beetles in 2002, and subsequently, it was used for marine sponges metagenomic data in 2004. Several putative Polyketide Synthases (PKSs) and Non-Ribosomal clusters coding for bioactive compounds were isolated and identified from *Theonella swinhoei* (Piel, 2002) using this same method. The sequence-based Screening method relies on PCR techniques, which either use primers or probes designed to hybridize with conserved regions known from the gene of interest. The sequence-based approach is flexible as it allows using heterologous strain for producing foreign genes and can accommodate the increase in the collection of phylogenetic markers. However, this method can only be used to identify new members of certainly known gene families only if there are regions of similarity to the probes or primer used. The limitation is that it undermines the detection of any new genes with unknown or unfamiliar sequence (Simon & Daniel, 2009; Singh et al., 2009)

### **2.2.2 Function-based Screening**

This screening method has been applied in many ways because it does not require having prior information about the gene or compounds and can detect new classes of full-length functional genes (Barone et al., 2014). In 2012, a Japanese group developed an efficient and fast functional screening process for detecting natural compound porphyrin by generating about 250,000 fosmid library using DNA from marine sponge *Discodermia calyx* (Barone et al., 2014). One of the most straightforward function screening processes is the use of indicator media containing a specific substrate, which, when it comes in contact with a surrounding enzyme gives a product which can be seen in the form of clearing zone around the transformants (Steele et al., 2009). However, the screening method involves the use of a compatible host that can express the transformed gene under the heterologous condition. Besides, more clones need to be analyzed than in the sequence-based approached to increase the accuracy of getting the positive clone (Simon & Daniel, 2009). Another interesting functional-based strategy is Substrate Induced Gene Expression Strategy System (SIGEX), which identifies a novel gene through a reporter gene that is co-expressed with the gene of interest. The promoter gene for a green fluorescent protein can be constructed and placed adjacent to the target gene in the expression vector so that the positive clones will be detected and separated by the fluorescent sorting techniques (Simon & Daniel, 2009).





**Figure (2): Steps involve in Next generation Sequencing Metagenomics approach, (1) various environmental samples site, 2) Extraction of RNA or DNA, 3)different sample preparation strategies, 4) sequencing technologies available 5) Different data analysis that could be done on the metagenomics data. Adapted from (D'Argenio & Salvatore, 2015).**

### 3. DNA Ligase

DNA ligase belongs to the family of enzymes called nucleotidyl transferases. This family is capable of transferring groups containing phosphorous group between biomolecules, a process that is essential in the metabolism of nucleic acids. DNA ligase is very crucial to all living cells because it plays vital roles in DNA replication, repair, and recombination (Pergolizzi et al., 2016; Timson, Singleton, & Wigley, 2000; Wilkinson et al., 2001). DNA ligases were first isolated from several sources by different research groups as early as the 1960s, and many ligases have also been generated in recombinant forms (Lehman, 1974). All DNA ligases contain essential lysine in the KXDG motif which is involved in its biochemical activity by transferring an adenylate group from a cofactor to the '5'phosphate of the DNA resulting in the sealing of breaks within the DNA strands (Pergolizzi et al., 2016; Wilkinson et al., 2001). DNA ligase can be classified as Adenosine 5' Phosphate ATP dependent type or NAD<sup>+</sup> Dependent type based on the co-factor or co-substrate present in the reaction site. ATP-dependent DNA ligases are common within Eukarya and Archaea, while NAD<sup>+</sup> Dependent DNA Ligase is exclusively found in many bacteria. Besides, some organisms such as *Bacillus subtilis*, *Mycobacterium tuberculosis*, *Neisseria meningitidis* and *Vibrio cholerae* may have more than one type of DNA ligase (Pergolizzi et al., 2016; Shuman & Lima, 2004; Wilkinson et al., 2001).

#### 3.1 Biochemical structure and functions of ATP Dependent DNA Ligases (ADLs)

This type of DNA ligase is found in all domains of life (Pergolizzi et al., 2016; Wilkinson et al., 2001). In many archaea studied, the only functional genes for DNA ligases encode the ADLs. However, in bacteria, NAD<sup>+</sup> dependent DNA ligases are most needed, and in some bacteria where ADLs are present, they always compliment the NDLs but not as a replacement (Korycka-Machala et al., 2007). ADLs in archaea exhibit homology in sequence and sizes compare to those found in bacteria that have a wide range of diversity in domain arrangement and total size. Although many bacteria do not possess the ADLs and the functions of bacteria's ADLs have not been fully described. Bacterial ADLs are assumed to be non-essential and are only expressed under specific growth conditions (Wilkinson et al., 2001). Four classes (Lig B, C, D, and E) of ADLs from bacteria are present, three of the four classes (Lig B, C, and D) descend from the same ancestor within the bacteria, and LigE ADL type was horizontally transferred to the bacteria separately (Williamson et al., 2016).

The Domain organization of DNA ligase varies between species (fig. 3). The different structural domain arrangements of ADLs can be found in the Pfam database (Williamson et al., 2016). All the ADLs have the core adenylation domain. Some of them, especially those that play an essential role in DNA metabolism, have extra domains that assist in DNA binding, linkage to other enzymes, or biomolecular complexes (Pergolizzi et al., 2016). The recent nomenclature of ligase domains is through specific Pfam identifiers on the Pfam database of proteins such as the adenylation domain which has the Pfam number PF01068 (Williamson et al., 2016)(Fig. 3). Some of the ligase domains are also identified with other proteins involved in DNA metabolism; for instance, PF0189 is recognized as DNA primase that synthesizes short RNA fragments. However,

some domains that are not identified with DNA LigD are usually annotated as Pol domain of LigD (Pergolizzi et al., 2016).

Furthermore, LigD also possesses a Nuclease domain having 3'-5' exonuclease activity that assists with proofreading during the ligation process; thus, ADLs exhibit greater diversity in their domain organization (Pergolizzi et al., 2016). LigD is very crucial to the Non-homologous end-joining process (NHEJ) that is involved in double-stranded breaks in DNA. It interacts with the Ku proteins involved in the NHEJ pathways; Ku proteins sometimes interact with Lig C when LigD is not present (Korycka-Machala et al., 2007; Shuman & Lima, 2004). Archaea ADLs are the major key player in the DNA replication process, unlike bacterial ADLs, which only function in DNA end-joining activities. However, the presence of bacterial ADL genes suggests that their protein products are integral to the DNA metabolism of the host cells (Pergolizzi et al., 2016; Shuman & Lima, 2004). Numerous functions have been suggested for bacterial ADLs, but these require further experimental validations (Korycka-Machala et al., 2007; Williamson et al., 2016)

### 3.2 Biochemical structure and functions of NAD<sup>+</sup> Dependent DNA Ligases (NDLs)

Most bacteria have NDLs genes, and this is critical for their efficient DNA replication in contrast to bacteria ADLs (Korycka-Machala et al., 2007). This essentiality makes them possible targets for antibacterial drugs and the production of novel necessary chemicals for biochemical characterization of DNA ligases (Pergolizzi et al., 2016). Although some archaeal organisms encode NDLs like *Methanomethylophilus alvus* and *iokiarchaeota*, which might have been gotten through horizontal gene transfer from bacteria as suggested through phylogenetic analysis studies (Spang et al., 2015; Zhao et al., 2006). Bacterial NDLs sequences show a high level of homology with highly conserved sequences relative to ADLs sequences, and the reaction mechanism is quite similar; only co-factors are different (Shuman & Lima, 2004; Wilkinson et al., 2001). NDLs, especially from bacteria, are structurally characterized to study their biochemical activities for developing possible inhibitors for DNA ligases (Pergolizzi et al., 2016).

The crystallographic structure of NDL reveals the adenylation domain of the enzyme from *Thermus filiformis*. Other high-resolution NDL structures have also been obtained from other classes of bacteria (Lee et al., 2000). The modular architecture of NDL has four domains, which are represented in the Pfam database, with some domains having subdomains (fig.3). Domain 1 (PF01653) has the subdomain 1a that binds beta NMN<sup>+</sup> moiety and the subdomain 1b (adenylation domain or nucleotidyltransferase domain) that binds to AMP moiety of NAD<sup>+</sup>. Domain 2 (PF03120) has an oligonucleotide binding site (OB). Domain 3 (3a-PF03119, 3b-PF14520) has both zinc finger and helix-hairpin-helix motif (HhH), and domain 4 (PF00533) is a BRCT domain which is at the C terminus.

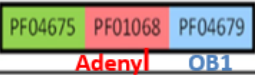
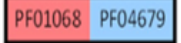
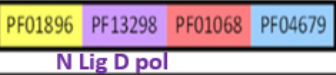

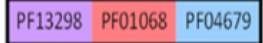


The adenylation domain (subdomain 1b) and OB-fold in domain 2 formed the catalytic core, which is flanked by N and C-terminal domains in a C shaped manner. The adenylation domain also has a subdomain 1b, which can interact with the  $\beta$  NMN moiety of  $\beta$ -NAD<sup>+</sup>, this subdomain is not found in the ATP dependent types. Subdomain 1a is important in beta-NAD<sup>+</sup>




binding and initiating the adenylation step but may not be needed in a case where the pre-adenylated ligase already binds the DNA. Crystallographic studies suggest that the adenylation subdomain1a closes when beta-NAD<sup>+</sup> is bound in a way to make the beta NMN side a leaving group and then exposes the adenosine phosphate lysine residue to a nucleophilic attack. The conformation of subdomain1a switches from open to closed-form by swiveling at least 180degree to accommodate this change. The adenylation subdomain 1b has some critical residues (KXDG, E173, K290, Y225, E113 in *E.coli*), which, when modified, can result in inactivation or a considerable decrease in the rate of enzyme activity.

The Oligonucleotide Binding site OB-fold (domain2) and helix-hairpin-helix motif, HhH (domain 3b), assist in clamping the DNA across 19bp. Domain 2 helps to cover the nicks and assist in fitting the double-stranded DNA into domain 1. Domain 3b binds the DNA in a non-specific manner along the minor groove. The subdomain 3a, a Cys4-type zinc finger-like, puts the OB and HhN at appropriate conformations during DNA binding. Domain 4, breast cancer carboxy-terminal, BRCT domain has not been investigated, but it seems not essential for the adenylation process. It is found as a flexible unit with no definite structure in the crystallography studies. In addition, DNA ligase without the BRCT domain still functions in a ligation reaction, although at a lower rate. Nonetheless, DNA binding was seen to be tighter in the presence of this domain, although there is no interaction directly with DNA. Besides, the BRCT domain could act as a signal transducer in a case of DNA damage, and it was also seen to enhance protein to protein interaction in the signaling process (Pergolizzi et al., 2016). Conclusively, a large conformational change occurs likely in many DNA ligases that catalyze ligation reactions irrespective of whether it is NAD<sup>+</sup> or ATP type.

### ATP Dependent DNA Ligase

Protein Family Name	Pfam Domain Architecture	Size of Proteins (number of amino acids)
LigB		860
LigC		512
LigD1		764
LigD2		868
LigD2		764
LigD3		419
LigE		499

### NAD<sup>+</sup> Dependent DNA Ligase

Protein Family Name	Pfam Domain Architecture	Size of Protein (number of amino acids)
<i>E. coli</i> LigA		671

**Adenyl OB ZBD HLH BRCT**

Figure (3): Conserved Pfam Domains within the NAD<sup>+</sup> and ATP Dependent DNA ligases from Bacteria and Archaea. Modified from Pergolizzi, G., *et al*, 2016.

### 3.3 Conserved Sequences in DNA ligase

Structural- functional analyses revealed that there are five conserved motifs in the primary sequences of DNA ligases, and they play a crucial role in the binding of nucleotide, nick identification, and nucleotidyl group transfer (Nishida et al., 2006). The catalytic core (adenylation domain and OB-fold domain) has active-site lysine in a residue called KXDG (Lys-Xaa-Asp-Gly), which is found in motif 1 (Wilkinson et al., 2001). The Glutamate (Glu) residue in motif III forms a hydrogen bond with the ribose residue of ATP, and the tyrosine residues with other essential residues help in stacking and stabilization of the active site and adenine conformation. Comparing the sequence homology within each ADLs and NDLs separately to the homology between both DNA types suggests that the two are very different in structure (Lim et al., 2001).

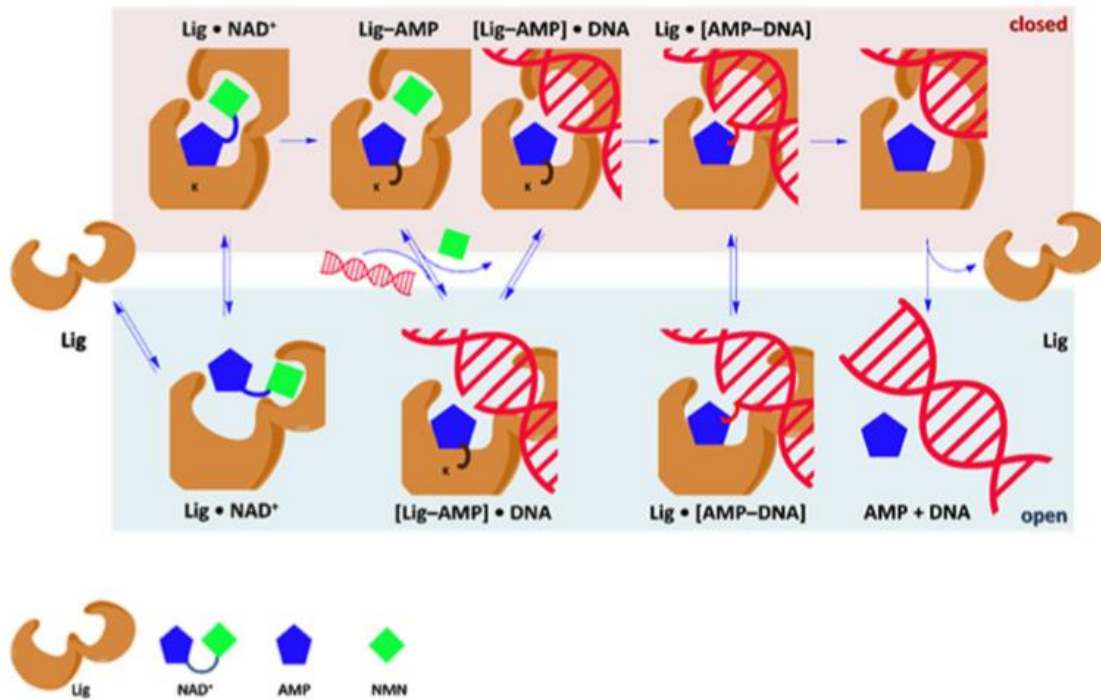


three domains of life. Another possible explanation is that the last ancestors of all the three domains (Bacteria, Archae and virus) initially carried the NAD<sup>+</sup> Dependent DNA ligase, but as ATP Dependent DNA ligase evolved, archaea and eukarya lost the NAD<sup>+</sup> type, but bacteria were somehow able to retain it. Later Bacteria were able to gain ATP Dependent DNA ligase from lateral gene transfer (Wilkinson et al., 2001). However, it was discovered that some ATP Dependent DNA ligases found in bacteria might not have an active function. For instance, *M. leprae* encodes an NAD<sup>+</sup> Dependent DNA ligase which is similar to putative ATP Dependent ligase from *M. tuberculosis*. However, unfortunately, the ATP type was found out to be pseudogenes with many stop codons (Cole et al., 2001).

### 3.5 Reaction Mechanism of DNA ligase

The study of LigA from *E. coli* has given significant insight into the reaction mechanism of many of NAD<sup>+</sup> Dependent DNA ligases (NDLs) since it is the most studied and the first bacterial DNA ligase ever to be biochemically characterized (Lehman, 1974; Shuman & Lima, 2004). NDLs ligation reaction uses a ping pong reaction mechanism that forms adenylylated intermediate (E-AMP) at the conserved nucleophilic lysine residue (KXDG) in the presence of  $\beta$ -NAD<sup>+</sup> (Pergolizzi et al., 2016). The binding of the cofactor,  $\beta$ -NAD<sup>+</sup>, leads to changes in the conformational state from open to a closed one (Lahiri et al., 2012; Pergolizzi et al., 2011). This process involves subdomain 1a (NMN) covering up on subdomain 1b (AMP), both of which are binding sites for  $\beta$ -NAD<sup>+</sup> moieties (Pergolizzi et al., 2011). DNA ligase is always activated when there is a transition from open intermediate to a close, and this results to multiple large-scale changes during the ligation reaction (Gajiwala & Pinko, 2004; Lahiri et al., 2012).

In LigA, the  $\beta$ -NAD<sup>+</sup> is initially recognized by the subdomain 1a which has the NMN binding site, and this process distinctively selects  $\beta$ -NAD<sup>+</sup> over ATP as a co-factor (Nandakumar et al., 2007). While in the closed conformation, the adenine of  $\beta$ -NAD<sup>+</sup> is bound tightly, and the adenine is held at a constant angle by interactions with several conserved residues (Nandakumar et al., 2007). However, the AMP moiety changes by rotating its ribose about the N-glycosidic bond, and alternating between syn and anti at various intermediate stages (Nandakumar et al., 2007; Pergolizzi et al., 2016). Similar large-scale conformational changes have been observed in many ATP dependent DNA ligase, but it varies slightly depending on the type of the ligase studied.



**Figure (5): Mechanism of Reaction for NAD<sup>+</sup> Dependent DNA ligases**

The N-terminal adenylation domain of Lig A has two subdomains (1a & 1b). The subdomain 1a includes the binding sites for NMN, and subdomain 1b contains the binding site for AMP. The enzyme transits from open to close conformation by binding of  $\beta$ -NAD<sup>+</sup> and subsequent closure of subdomain 1a on to subdomain 1b. The ligation reaction involves a ping pong mechanism, which is a non-sequential mechanism that makes the binding of a first substrate to cause the enzyme to change to an intermediate form that allows the binding of the second substrate. In NAD<sup>+</sup> DNA Dependent reaction, there is always a formation of closed intermediate after an open intermediate before the next ligation reaction. While the adenine ring is always held tightly by several conserved residues, the AMP moiety changes from syn in non-covalently bound NAD<sup>+</sup> to be anti in adenylate-ligate intermediate and then back to syn in the adenylated-DNA intermediate. Adapted from Pergolizzi *et al.*, 2016.

### 3.6 Nick Determination and Discrimination in DNA ligases

A DNA nick has 5'-phosphorylated terminal side and 3'-OH on the opposite terminal side in a phosphodiester break. It has been shown that 5' phosphate is necessary for the nick recognition while the 3' is required for the DNA binding, but in a case where the nick does not have 5' phosphate terminal end, there is a drastic reduction in the formation of the nick-specific complex (Shuman & Lima, 2004). Ligases are also capable of discriminating between nicks and gaps; for



instance, a more than one nucleotide gap in a nick site can eliminate the DNA binding (Doherty & Suh, 2000; Odell & Shuman, 1999). The footprinting studies in ATP dependent DNA ligases from T7 and Chorella virus discovered that ligases bind asymmetrically to the nicks, and they extend 3-8 nucleotides on the 3'-OH side and 7-12 nucleotides on the 5' phosphate sides of the nick (Odell & Shuman, 1999). The DNA binding patterns reveal that the DNA is docked in a positively charged cleft region, which lies between the first two domains of DNA ligase. It has also been shown that the interdomain linker region (motif V) of a PCBV-1 DNA ligase bound to nicked DNA is less accessible to proteolytic attack (Doherty & Suh, 2000; Odell & Shuman, 1999). In addition, mutagenesis and photocrosslinking studies have suggested that two conserved lysine residues in motif V are essential for DNA nick recognition (Doherty & Suh, 2000).

Domain 1 and 2 of T7 ligases have been reported to have unspecific DNA binding activities, and through nick sensing, ligases were able to perform distinct DNA binding. These two domains of DNA ligase were said to be important for nick sensing and ligation activities (Doherty & Suh, 2000). Ligase must be adenylated at the active site lysine for a nick to be recognized (Shuman & Lima, 2004). In the non-adenylated state, the conformation of ATP type T7 ligase is such that the DNA binding site of the OB domain is rotated away from the active site cleft while in NAD<sup>+</sup> type Tfi ligase, the OB domain is facing the active site cleft. In non-adenylated T7 ligase structure, the OB domain is oriented in such manner to prevent DNA binding before adenylation at the conserved active site lysine residue (Doherty & Suh, 2000). The OB domain of NAD<sup>+</sup> Dependent DNA ligase performs a double role during the ligation process. Firstly, it rotates its position to prevent an attack on ATP from incoming nucleophiles, thus, enhancing adenylation. This adenylation then acts as a conformational switch that brings the DNA binding site closer towards the active site cleft so that only adenylated ligase will bind to the nicked DNA (Doherty & Suh, 2000). Larger DNA ligases have additional DNA binding motifs like zinc finger domain and HhH motifs that increase the DNA binding surface (Doherty & Suh, 2000).

### 3.7 DNA Ligase Assays

Assays involving DNA joining can be easily detected using various biochemical and biophysical techniques (Bowater et al., 2015). One of the most widely and convenient used techniques in primary biochemical studies of DNA ligase is Gel electrophoresis. Although gel electrophoresis is very common, it may not be the ideal method for drug discovery study because it is intensive and time-consuming (Pergolizzi et al., 2016). Some alternatives methods have been devised for biochemical analysis involving the study of drugs and inhibitors. Some of these methods rely on novel forms of some important derivatives such as designed fluorescent beta-NAD<sup>+</sup>, which can act as co-factor for NAD<sup>+</sup> type DNA ligases and many other beta-NAD<sup>+</sup> types consuming enzymes (Lohman et al., 2016). Fluorescently labeled probes are increasingly used in biophysical and chemical studies to detect the joining of DNA strand breaks (Pergolizzi et al., 2011)

Other approaches include fluorescence quenching, molecular beacon-based methods, electrochemical methods, nanoparticle-based sensor, and surface plasmon resonance (Pergolizzi et al., 2016). Some biochemical DNA ligase assays do not involve labeling of the nucleic acid, and this includes the use of nicked plasmid DNA substrate and mercury-based electrodes (label-free electrochemical approaches) (Vacek et al., 2008). The combination of different techniques has made DNA ligase assay quicker and more effective; for instance, the incorporation of digoxigenin-labeled probes in a double-surface enzyme-linked electrochemical assay (Stejskalová et al., 2014).

Furthermore, oligonucleotide templates forming stem-loop structures immobilized on different surfaces by different linkage to gold surfaces or biotin-streptavidin interactions are standard with many recent DNA ligase assays (Pergolizzi et al., 2016). The presence or absence of these labels makes it easier to detect the changes in the DNA backbone structure. In a case where a nick is included in the DNA stem-loop structures, the release of a signal which can be monitored through fluorescence or electrochemical techniques makes ligation reaction easily detected (Scott et al., 2006). In addition, these techniques involving appropriate substrates and assays have made the study of the interaction between DNA ligases with other enzymes such as nucleases and others involve in nucleic acid metabolism possible (Scott et al., 2006).

### **3.8 Industrial and Biotechnological Applications of DNA ligase**

#### **3.8.1 DNA ligase applications in gene cloning**

DNA ligase has been critically useful for many years in gene cloning and mutation detection involving ligation chain reaction (Barany, 1991; Wilson et al., 2013). One of the enzymes that have evolved to be most important to molecular biologists is T4 DNA ligase, an ATP dependent enzyme from bacteriophage T4 (Wilson et al., 2013). It is the first DNA ligase to be discovered, and it is responsible for sealing Okazaki fragments during phage genome replication (E. S. Miller et al., 2003; Weiss & Rich, 1967). This physiological role of T4 DNA ligase has made it essential *in vitro* as a highly efficient DNA ligase for sealing single-stranded nicks in duplex DNA (Wilson et al., 2013). Also, T4 DNA ligase can ligate double-stranded DNA (ds DNA) fragments that have overhangs, single-stranded (cohesive) or complementary ends, and these techniques together with blunt end ligation are highly required molecular biology protocols (Barany, 1991; Wilson et al., 2013). Besides, it is one of the commercially available ligases that can join blunt-ended DNA duplexes without the help of enhancers such as polyethylene glycol (Sgaramella et al., 1970).

#### **3.8.2 DNA ligase as a molecular engineering tool in DNA assembly**

Taq DNA ligase (Taq lig) is among the enzymes in DNA assembly that can be used for constructing natural and synthetic genes, genetic pathways, and even whole genome (Gibson, 2011; Wilson et al., 2013). Type II restriction enzymes are commonly used for cloning DNA

fragments by inserting them into a vector (Yount et al., 2002). However, this technique has limited uses in DNA assembly because it involves simultaneously assembling of multiple DNA pieces that do not have restriction sites (Gibson, 2011). Numerous *in vitro* enzymatic strategies have been developed to assemble single-stranded (ss) oligos into large constructs of double-stranded(ds) DNA, which can further be cloned, for instance, the polymerase cycling assembly couple to PCR amplification(Gibson, 2011). Three enzymes are involved in the assembling of overlapping DNA molecules, and they include an exonuclease, polymerase, and ligase. These three enzymes participate in three different DNA assembly methods. The first method utilizes Taq DNA polymerase, 3' exonuclease of T4 DNA polymerase (T4 pol) and Taq DNA ligase (Taq lig) in a two-step thermocycled reaction. The second method employs 3' exonuclease III, antibody-bound Taq pol, and Taq ligase in a one-step thermocycled reaction. The third method, which is the simplest one, is also a one-step thermocycled reaction (ISO) that integrates 5'-T5 exonuclease, Phusion DNA polymerase, and Taq ligase to assemble both ssDNA and dsDNA (Gibson, 2011). It is easier to work with the ISO system because all the components can be premixed and stored for further use. It also has the advantage of generating linear assemblies when the exonuclease is inactivated during the reaction (Gibson, 2011). DNA ligase has also been applied to gene synthesis involving circular assembly amplification. It assists in the ligation of the oligonucleotides immediately after the construction of exonuclease-resistant circular DNA(Bang & Church, 2008).

### **3.8.3 Use of ligase in Next- Generation Sequencing (NGS) technologies.**

Ligases are useful as ligation adapter during the preparation of sample in many next-generation sequencing methods( Illumina, 454, Ion Torrent sequencing), for instance, in Illumina, RNA fragments are ligated to pre-adenylated "3' adapter and "5' adapter using T4 RNA ligase 2 (Rnl2) which is subsequently followed by reverse transcription and PCR amplification(Van Dijk et al., 2014; Wilson et al., 2013). Ligase is also involved in sequencing reactions in SOLID sequencing methods(Wilson et al., 2013). In SOLID systems, ligase is used to seal the nick between the "5' end of the growing strand and the "3' end of the oligo(Huang et al., 2012). The major difference between sequencing by ligation and sequencing by synthesis is that the former uses ligase to seal the nick in between elongating and complementary strands while the latter uses DNA polymerase for the same purpose(Huang et al., 2012).

### **3.8.4 DNA ligase' uses in DNA Origami**

A novel application of DNA ligase in nano chemistry is in organizing nanomaterial with reliable precision. DNA lattice has found wide applications in nanoelectronics, photonics, and biomolecules. Assembling of DNA nanostructures requires many strategies, which include an enzyme-free strategy that was used in assembling Y shaped bDNA structures on aluminum foil cationic surfaces during periodic 2-dimensional (2D) bDNA lattice preparation. Although, without the ligase, the small overhangs can easily assemble on the aluminum foil, however, DNA ligase



can speed up the making of the DNA lattice structure using the DNA overhangs (Bhanjadeo et al., 2017).

### 3.8.5 DNA ligase in Molecular Diagnostics

Various nucleic acid-based diagnostic methods for genetic disorders, bacterial, viral, and other pathogens have emerged for years due to the advent of PCR (Erlich et al., 1991). In later years, many other DNA amplification methods such as ligase chain reaction (LCR) and Q-beta replicase were developed as a complement or alternative to PCR (Wolcott, 1992). LCR has stood out as an effective diagnostic technique that has been commonly used to complement PCR amplification due to its ease of compatibility and ability to accurately discriminate DNA sequences with just one single base pair difference (Barany, 1991). This has further broadened the scope of single nucleotide polymorphism detection (SNPs), a method that identifies genetic diseases and infectious agents through insertions, deletions, or substitutions in genes (Barany, 1991). LCR has the advantage of limiting false-positive ligation products and improving sensitivity because it uses thermostable ligase. Previous ligation amplification reactions relied on the fresh addition of mesophilic DNA ligase after every denaturation step (Wu & Wallace, 1989).

LCR relies on a system that uses already synthesized two adjacent oligonucleotide primers, which are ligated and specifically hybridized to one strand of the target DNA (Barany, 1991). The junction between the two primers is usually arranged to place the nucleotide at the 3' end of upstream primer to where the single nucleotide difference in the target DNA is. The complement between the target sequence and the 3' end of the upstream primer at this site will result in the action of ligase joining them covalently, thus, generating a ligated product or otherwise a SNPs has been detected. The ligated product can then serve as a template for the next reaction in the presence of a thermostable ligase, similar to a cycling reaction in PCR amplification. In case a mismatch is detected at the primer junction, there will be no ligation due to discrimination in place by the thermostable ligase (Wiedmann et al., 1994).

Many modifications have been done to LCR; for instance, there is a ligase detection reaction, LDR, that is similar to LCR except that it performs a linear amplification (Landegren, 1993). Numerous mutations can easily be detected and analyzed with a single amplification when LDR is coupled with PCR, and this has been used in simultaneously detecting multiple mutations in cystic fibrosis and 21-hydroxylase deficiency (Eggerding et al., 1993). Another modified LCR with improved specificity is pLCR, which was designed to include a gap between the discriminating and non-discriminating primer. In this gap, there is Taq polymerase Stoffel fragment, which elongates the discriminating primer after ligation reaction (Landegren, 1993; Wiedmann et al., 1994). Quantitative assays for detecting point mutation have also been developed based on gap ligation reaction, and it has been useful in uridine to cytidine RNA editing events in WT-1 Wilm's tumor suppressor (Stewart et al., 1998).

Several Quantitative real-time LCR-based methods have also been developed for quantifying SNP in pools of DNA for a routine high throughput diagnostics (Psifidi et al., 2011). In addition, the colorimetric detection systems have been coupled to thermostable high-fidelity DNA ligase for the detection of point mutations in many gene-mutant diseases (J. Li et al., 2005). As technology advances, ultrasensitive detection system combining chemical ligation chain reaction and nanoparticles are rapidly developing (Shen et al., 2012). Recently, a cross-linked gold nanoparticles based assay system which combines the enzyme-free LCR, dispersed gold nanoparticles and streptavidin-modified magnetic beads were developed for detecting mismatch in RNA and DNA with high discrimination. This assay specifically discover the target DNA or RNA from the cell lysate (Kato & Oishi, 2014). LCR has been upgraded further to be a powerful genotyping assay by integrating several detection strategies, including various biosensors, DNAzyme, quantum dots (QD), surface-enhanced Raman scattering (SERS) and fluorescence resonance energy transfer (FRET) (Gibriel & Adel, 2017).

### **3.8.6 NAD<sup>+</sup> Dependent DNA ligases as chemotherapeutic targets in determining novel antibiotics candidates.**

Previous attempts have been made to inhibit DNA ligase by using modified derivatives of ATP or NAD<sup>+</sup>, but the likelihood to have non-substrate activities have limited the chances of any inhibitor targeted directly to the enzyme (Pergolizzi et al., 2011; Shuman, 2009; Wilkinson et al., 2001). Many different types of modifications have produced a wide range of effects, but it is difficult to understand the mechanism of those effects (He et al., 2011; Pergolizzi et al., 2011). The first set of DNA ligase inhibitors which has a promising effect for drug discovery and therapeutics were discovered through screening of libraries of derivatives of compounds like alkaloids, flavonoids, quinoline and quinacrine and generating suitable drug-like analogs that are more selective, show improved activity with high solubility (Dwivedi et al., 2008; Tripathi et al., 2011). One of the biggest factors that make the design of DNA ligase inhibitor difficult is the large conformational changes they undergo (Dwivedi et al., 2008; Gu et al., 2012; Pergolizzi et al., 2016). However, crystallographic studies and structures modeling of many NAD<sup>+</sup> Dependent DNA ligases with the virtual screening of molecular interactions have given valuable insights on potential inhibitors. Although, there is still need to shed more light on the complex molecular dynamics of the enzyme which would better facilitate the design of structure-based inhibitors (Dwivedi et al., 2008; Nandakumar et al., 2007; Tripathi et al., 2011; Yadav et al., 2015). One of the past accomplishments in resolving the structures of NAD<sup>+</sup> Dependent DNA ligase is the discovery of a hydrophobic tunnel which lies between the N-1, C-2 and N-3 of the adenine (Pergolizzi et al., 2011). This has further widened the opportunity of developing inhibitors that are selective and specific towards the NAD<sup>+</sup> type only since the ATP Dependent DNA ligases do not have this hydrophobic tunnel (Pergolizzi et al., 2016). Many potential inhibitors have also been examined, and some have been modeled and synthesized to have the same orientation of adenine, ribose and C-2 substituents in relation to DNA ligase active site. Some of the inhibitors that have comparable inhibitory activities include 6-azaindazoles, 2-amino-(1,8)-naphthyridine-3-carboxamides (ANCs) and aminoalkoxy pyrimidine carboxamide (AAPCs) (Pergolizzi et al.,

2016). Other tested inhibitors include phosphorylated derivatives such as AMP, cAMP, ADP and ATP replaced on C-2 of the adenine ring have also shown promising results. The limitations faced in generating NAD<sup>+</sup> inhibitors include difficulty in synthesizing the  $\beta$ -NAD<sup>+</sup> derivatives, weak stability and unspecific binding as a result of flexible functional structures (Pergolizzi et al., 2016).

#### **4. Rigidity, flexibility and enzyme thermostability**

Thermostable enzymes are of utmost importance in industrial and Biotechnological sectors. One of the major goals in protein engineering is to design a protein with improved thermostability (Eijsink et al., 2004). Over the years, a significant attention was focused on identifying and optimizing the weak spot regions in proteins to enhance their thermostabilities (Yu et al., 2015). Flexible sites have been indicated to be the weak spots in many proteins, and attempts have been made in many previous studies to identify and rigidify these sites (Yu & Huang, 2014). Many structural studies of mesophilic proteins have revealed that their thermostabilities can be improved by increasing the rigidity in their structures (Yu & Huang, 2014). The relationship between flexibility, stability and activity of proteins has been debatable for years (Yu et al., 2015; Yu & Huang, 2014). In thermophilic proteins, rigidity is more important to maintain the native folded structure of these proteins at high temperatures, but a certain level of flexibility is also needed to maintain the activities of these proteins. For many other proteins, flexibility is more important to perform functions such as ligand binding, macromolecular interactions and activity modulations in their native folded states (Yu et al., 2015). The 3D structures of many proteins showed varying degree of flexibility with some exhibiting lesser mobility. Other proteins have large disordered regions difficult to be resolved by techniques such as X-ray crystallography (Smith et al., 2003).

Generally, some regions in the secondary structures of proteins have been observed to show a greater degree of flexibility and not part of the regular structure (Dunker et al., 2001; Wright & Dyson, 1999). These unstructured parts are often referred to as Intrinsically Disordered regions (IDR). Intrinsically Disordered Proteins (IDP) contain large disordered regions or have completely disordered domains. On the other hand, globular domains are ordered regions that always have regular secondary structures (Linding, Jensen, et al., 2003). Functional regions in proteins can be found in both the globular and disordered regions (Linding, Jensen, et al., 2003). Many computational tools, such as PONDR and DisEMBL have been used to distinguish the disordered regions from the globular regions (Linding, Jensen, et al., 2003). Protein disorders help in understanding the protein function and its folding pathways. It has been implicated in protein misfolding and aggregation that leads to various diseases (Bates, 2003; Korhonen & Lindholm, 2004).

IDRs are present in both the eukaryotes and prokaryotes proteomes where they also play essential roles in the organisms, for instance, fibronectin-binding protein A and prokaryotic ubiquitin-like protein in bacteria have long IDRs (Dunker et al., 2001; Iakoucheva et al., 2002). In

eukaryotes, IDPs carry out functions relating to signal transduction and transcription regulation (Iakoucheva et al., 2002). IDRs/IDPs have been suggested to be heat resistant up to boiling temperatures as reported from their previous purifications and proteomic identification (Burra et al., 2010; Tompa, 2002). IDPs have also been reported to be cold-resistant; some disordered plant dehydrins have been implicated in the plant adaptation to freezing temperatures (Burra et al., 2010; Kovacs et al., 2008). Some researchers have suggested that IDPs may be part of the evolutionary strategy used by extremophiles for adaptations (Burra et al., 2010). However, since the decrease in protein flexibility increases protein thermostability and IDRs coincides with flexible regions, there is a possibility that they came from a different evolutionary line (Thompson & Eisenberg, 1999). In a previous structural comparison study, hyperthermophilic and psychrophilic organisms were reported to have a low to average protein disorder, but variability in protein disorders was reported for mesophiles and thermophiles (Burra et al., 2010). The relationship between IDPs and thermal stability is debatable, but the logic is that increased protein thermostability can be achieved by reducing flexible loops (loops with random mobility) that cause unfolding and fusing terminal IDPs. This can reduce irreversible aggregation (Sharma et al., 2009; Thompson & Eisenberg, 1999).

### **5. Role of Proline and Arginine in conferring thermostability to proteins**

Proline is a unique amino acid that has been associated with improved thermostability in thermophilic proteins in many previous studies (Barzegar et al., 2009; Watanabe et al., 1991). The proline residue is known to improve rigidity and decrease flexibility around the protein environment (Georlette et al., 2000). Proline is an exceptional amino acid in protein thermostability because of two reasons, firstly, it has a pyrrolidine ring which imposes rigid constraints ( $\phi$ ) on the N-C rotations, restricts the conformational space of surrounding residues and disrupts the tertiary conformation of protein through its imino acid (Igarashi et al., 1999; Watanabe et al., 1991). Secondly, in the unfolded state, it enhances protein thermostability by decreasing the conformational entropy (Barzegar et al., 2009). A single proline substitution has been reported to enhance thermostability in many enzymes such as glucosidase, alcohol dehydrogenase, amylase acidic xylanase, L-asparaginase and phosphoglycerate kinase (Bailey et al., 1990; Goihberg et al., 2007; Igarashi et al., 1999; L.-Z. Li et al., 2007; Yang et al., 2017; C. Zhou et al., 2010)

Arginine has also been implicated to be a stabilizing residue for many thermophilic proteins (Sandeep Kumar et al., 2000). Arginine has been reported to dominate the exposed surfaces of thermophilic proteins (Yokota et al., 2006). Arginine is exceptional compared to other positively charged/basic amino acids because of its guanidinium group that singly contributes three nitrogen atoms to form ionic interactions in three possible directions (Sokalingam et al., 2012; Strickler et al., 2006). Arginine has been reported to form more salt bridges and hydrogen bonds than other basic amino acids due to its larger electrostatic interactions (Meuzelaar et al., 2016). Increased salt bridges and hydrogen bonds have been widely suggested as strong indicators for thermal stability (Elbehery et al., 2017; Meuzelaar et al., 2016). It has been reported that arginine/glutamate

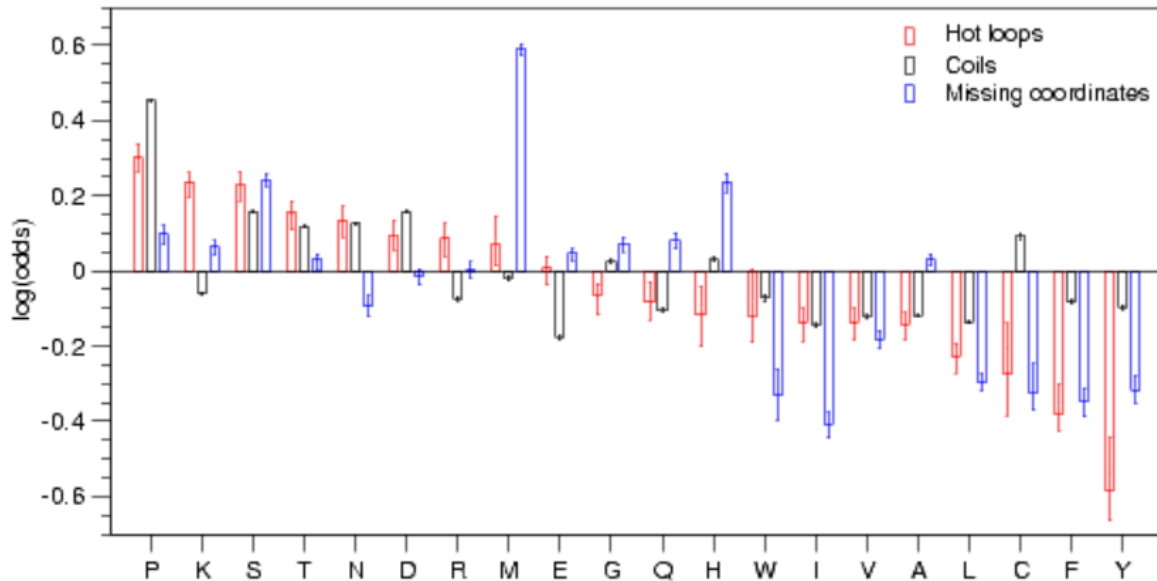
pairs form salt bridges with more favorable geometry than salt bridges formed in lysine/glutamate pairs (Meuzelaar et al., 2016). Mutating a surface lysine to arginine has been reported to significantly improve protein thermostability in green fluorescent protein (Sokalingam et al., 2012).

### **5.1 Relationship between proline and arginine residues in the disordered regions to thermostability**

IDRs/IDPs are biased in their amino acid compositions. IDPs are abundant in disorder-promoting amino acid residues (Proline(P), Arginine(R), Glycine(G), Glutamine (Q), Glutamate (D), Serine (S) and Lysine (K)) and depleted in order-promoting amino acid residues (Tryptophan(W), Cysteine(C), Phenylalanine(F), Isoleucine(I), Tyrosine(Y), Valine(V) and Leucine(L) (Dunker et al., 2001). Although there is no unified definition of protein disorder, each definition gives a different perspective. Thermodynamically, protein disorder in a polypeptide chain is when the protein is in a random coil structural state where it exhibits a maximum degree of freedom within the conformational space.

In DisEMBL server, several definitions of protein disorder are integrated and used to predict disorders in proteins using artificial neural networks. The server defines protein disorder as two-state models where each residue is either ordered or disordered, and each residue can be disordered by loops, hot loops or missing coordinates (REMARK465). The loops may not be disordered, but within loops there may be protein disorder. The region of loops with a high degree of mobility as defined by C- $\alpha$  temperature factors (B-factors), is called hot loops. The B factors was used to derive the propensity scales of each residue from being disordered. According to DisEMBL, the propensity of amino acids to be disordered is preferably defined by the hot loops (Linding, Jensen, et al., 2003).

Structural comparisons between mesophilic and thermophilic protein have revealed that more proline residues in the hot loops give more thermostability to the latter (Barzegar et al., 2009; Bogin et al., 1998; Sakaguchi et al., 2007). The mechanism of protein thermostability that has been widely reported is the high frequency of proline within the loop region (Barzegar et al., 2009). In addition to the high frequency of proline in the hot loops, the position of prolines in the 3D structure plays critical roles in the protein thermostability (Barzegar et al., 2009). In alcohol dehydrogenase, prolines inserted at the external loops significantly increased protein thermostability, whereas the ones inserted outside the loop regions only moderately increased the thermostability (Bogin et al., 1998). The proline residues at these regions constrained the loop structures by increasing the rigidity and decreasing the flexibility, a common mechanism of maintaining protein thermostability (Barzegar et al., 2009; Watanabe et al., 1991).



**Figure (6): Propensities of amino acids to be disordered according to Russell/ Linding hot loops and loops definitions.** The scale represents the data used by the author(Linding, Russell, et al., 2003) which was generated by neural networks algorithm in predicting the disorder. Error bars are to 25 and 75 percentiles, as estimated by stochastic simulation. Adapted from Jensen, et al., 2003 (Linding, Jensen, et al., 2003)

## 6. Rationale and goal of the study

The increasing demand for novel enzymes with unique properties such as heavy metal resistance and ability to survive a broad range of conditions such as temperature, pH, and salinity have made biocatalysts from extremophiles incredible assets to the biotechnology industry (Lorenz & Eck, 2005). For instance, the Red Sea Atlantis II brine pool LCL is characterized by a temperature of 68.2 °C and salinity, which is 7.5 times that of normal seawater (Karbe, 1987; A. R. Miller et al., 1966). These characteristics make the extremozymes from such an environment suitable for industrial biotechnological use. The advent of metagenomics has made the exploration of many novel biocatalysts from these several unculturable microorganisms possible (Uchiyama & Miyazaki, 2009). DNA ligase is an invaluable enzyme in the biotechnology industry, and it is apt to say that such a source of enzymes like Red Sea brine pools will indeed confer the unique enzymatic properties.

The goals of the study include:

- 1) To screen for putative DNA ligases sequences in the metagenomes of the Red Sea Brine pools (Atlantis II, Discovery Deep Brine pools and Kebrit Deep interface layer) using bioinformatic analysis, and identify at least a potential DNA ligase under ATP-Dependent DNA ligase and NAD<sup>+</sup> Dependent DNA ligase for synthesis.
- 2) To apply bioinformatics tools to primarily compare the thermostability of ligase sequences across metagenomic assemblies of different extreme environments, including the Red sea brine environments.
- 3) To express and purify at least one of the synthesized DNA ligase.



## Chapter 2: Materials and Methods

### 1. Sample Collection

Samples containing 100 liters of water each, were collected from the lower convective layer of Atlantis II brine pool and Discovery Deep brine pools subsurface sediment. 120 liters of water sample was also collected for Kebrit Deep Upper and Lower Interfaces. The Atlantis II and Discovery Deep samples were collected from the depth close to 2000 m, and that of Kebrit upper and lower interfaces samples were gotten from the depth very close to 1500m. All the samples of water were collected from each depth by the aid of shipboard Niskin bottles that were connected to CTDs (conductivity, temperature, and depth sensors). In addition to the CTDs, the seabird DO sensor was also linked to measuring oxygen saturation. Serial filtration using Mixed Cellulose Ester (Millipore) filters of different pore sizes of 3, 0.8, and 0.1 $\mu$ m was done, and that of 0.1  $\mu$ m processed filters were further collected, preserved in sucrose lysis buffer and stored at – 20°C freezer. They were later transported to the American University in Cairo laboratory where they were stored at – 80°C freezer before DNA extraction. These previous procedures were done by the KAUST Red Sea R/V Aegaeo expedition team in spring 2010 (Abdallah et al., 2014; Sayed et al., 2014; Siam et al., 2012).

### 2. DNA Extraction, Sequencing, and generation of Metagenomic Libraries

The extraction of total genomic DNA retained on each treated 0.1 $\mu$ m filter applied to the four different samples was done using the Metagenomic DNA Isolation Kit for Water (EPICENTRE, Biotechnologies, Madison, WI, USA). The concentration of the DNA was determined by PicoGreen assay using NANODROP™ 3300 Fluorospectrometer (Thermo scientific, USA), and the quantities of this DNA were sufficient for Direct sequencing. Following the GS FLX Roche Titanium library guide, the metagenomic libraries of all the four different interface layers were constructed. The Double SPRI method (Hawkins et al., 1994) was chosen for selecting the DNA fragment sizes. Pyrosequencing of the metagenomes was done using 454 GS FLX Titanium technology (454 life sciences). The quality of the metagenomic reads was checked by PRINSEQ-lite v0.204 (Schmieder & Edwards, 2011) and CD-HIT- 454 (Niu et al., 2010). The metagenomic datasets of each brine pools were established and made available on the NCBI. All these previous steps were done at the Biology Department, The American University in Cairo (Abdallah et al., 2014; Sayed et al., 2014; Siam et al., 2012)

### 3. Bioinformatics Analysis I

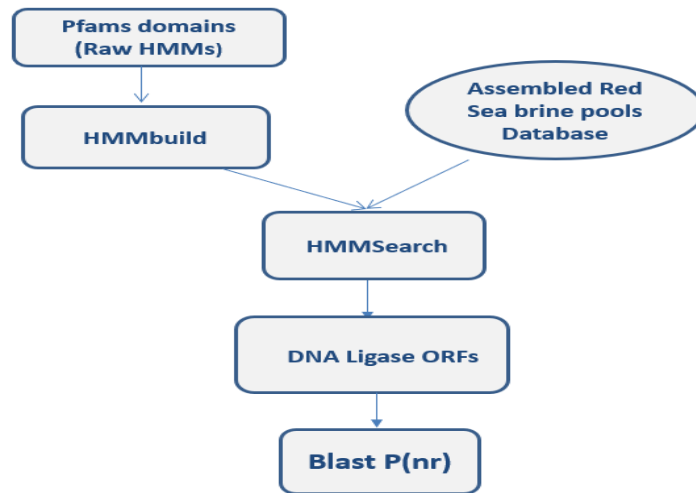
#### 3.1 Contig Assembly

The metagenomic datasets of each Red Sea brine pools (Atlantis II, Discovery Deep, Kebrit Deep Upper, and Lower interfaces) were downloaded from the Sequence Read Archives (SRAs), National Center of Biotechnology Institute (NCBI). Each sample was assembled into contigs using MEGAHIT assembler v 1.0 (D. Li et al., 2016). ORFs were predicted from the assembled final contigs using Metagene Annotator (MGA), using default parameter (Noguchi et al., 2008).



### 3.2 DNA Ligase Screening from the Red Sea brine pools Metagenomic datasets

The Pfam accessions for all the DNA ligase domains, including the NAD<sup>+</sup> and ATP types, were obtained from the Pfam database 32.0. The domains selected were seven domains in total, three domains for the ATP type; N terminus ( PFO4679), C terminus (PFO4679) and ATP dependent DNA ligase ( PFO1068) and four domains for the NAD<sup>+</sup> type; Adenylation (PFO1653), Oligonucleotide Binding, OB (PF14743), Zinc fingerlike Binding, ZBD (PFO3119) and BRCA 1 C terminal, BRCT (PF0533). The HMMs of the seven ligase domains were downloaded from the Pfam database 32.0 to build a concatenated HMM model. This was further used in HMM search against the assembled ORFs of the final contigs. The putative DNA ligase ORFs from the final HMM output were assessed based on the length (>100), E values (1e-5), and the coverage of the domains.



**Figure (7): The Pipeline for screening DNA ligases from the assembled Red Sea Metagenomic datasets.** The 7 DNA ligase domains selected include (N terminus ( PFO4679), C terminus (PFO4679) and ATP dependent DNA ligase ( PFO1068), Adenylation (PFO1653), Oligonucleotide Binding, OB (PF14743), Zinc fingerlike Binding, ZBD (PFO3119) and BRCA 1 C terminal, BRCT (PF0533).

### 3.3 Multiple Sequence Alignment (MSA) of all the Selected DNA Ligase ORFs from Red Sea Metagenomic Datasets

All the Red Sea DNA Ligase sequences were aligned using the Clustal Omega (<https://www.ebi.ac.uk/Tools/msa/clustalo/>) to identify the essential, conserved, and active site motifs in the DNA Ligase sequences. The two synthesized ORFs (LigATL1 and LigKDU4) were further aligned to Bacterial DNA ligase sequences. The MSA file was visualized, colored, and edited using Jalview.

### **3.4 Phylogenetic Analysis of Red Sea Metagenome Putative Lig ATL1 (ATP type) and LigKDU4 (NAD<sup>+</sup> type).**

The phylogenetic tree was inferred using the Maximum likelihood method with bootstrap testing (1000 replicates) and the Jones-Taylor-Thornton (JTT) model as the substitution model. The retrieved DNA ligase sequences of both the ATP and NAD<sup>+</sup> types from the NCBI with the Red Sea DNA ligase sequences were aligned using the MUSCLE Algorithm in MegaX, and the phylogeny tree was also generated in MegaX (Kumar et al., 2018).

### **3.5 3D Modelling and Predicting the Superimposition of LigATL1 (ATP type) and LigKDU4 (NAD<sup>+</sup> type)**

Protein Fold Recognition server (PHYRE 2), set at the default parameter, was used in comparative homology modeling of the proteins using a specific DNA ligase template (Kelley & Sternberg, 2015). The superimposition of both proteins was also predicted from the same server using the best alignments.

### **3.6 Prediction of Molecular weight and Isoelectric point(pI) of LigATL1 and ligKDU4**

The theoretical pIs and the molecular weight of both ligase genes were predicted from isoelectric.org and pI/Mw analysis tool of the Expasy server, set at default.

## **4. Bioinformatics Analysis II: Comparative thermostability analysis of Red Sea Ligase sequences to ligase sequences of different environments; stability roles of Proline and Arginine**

### **4.1 Retrieving different ligase sequences from the NCBI**

The ligase sequences from different metagenomic assemblies were retrieved from NCBI (<http://www.ncbi.nlm.nih.gov/>). This was done by searching the ligase sequences found under the species/ genus that have the best hits to Red Sea putative ligase sequences in the NCBI database. The ligase sequences from different bio projects were selected based on the environment temperatures.

The ligase sequences from the Kebrit deep interface layers and Atlantis II, LCL were used as the models for the mesophilic environment and thermophilic environment, respectively. For example, the ligKDU4 (MLK) depicts the DNA ligase from the mesophilic Kebrit upper surface environment which has a temperature of 23.3<sup>0</sup>C (Winckler et al., 2001). The other two ligase sequences selected were from NCBI search of different bio projects found under *Candidatus* Marinimicrobia bacterium which is the closest taxonomic group having best hit of ligKDU4. HLG represents the ligase sequence from the hyper-thermophilic Guayamas base sediment metagenomic assembly (Seitz et al., 2019), while PLC represents ligase sequence from psychrophilic cold, oxic subseafloor aquifer metagenomic assembly (Tully, Wheat, et al., 2018).

**Table 1: Environments and abbreviations of retrieved ligase sequences used in the comparative thermostability study**

The environments of the selected ligase sequences	Temperature (°C)	Abbreviation
Hydrothermal Guayamas base sediment (Seitz et al., 2019)	>120	HLG
Hot spring Iceland (Clum et al., 2009)	95	HLH
Atlantis II, Lower Convective Layer (LCL)	68	TLA
Yellowstone National park(mineral samples)((Tatusova et al., 2014)	43	TLY
Kebrit Deep upper surface interface	23.3	MLK
Terrestrial habitat (root nodules of legumes) (Mantelin et al., 2006)	28	MLT
Cold, oxic sub sea floor(Tully, Wheat, et al., 2018)	4	PLC
Russian lake Baikal(sub ice water) and japan lake Biwa(Cabello-Yeves et al., 2018; Hiraoka et al., 2019)	1-10	PLB
Mediterranean sea (marine plankton)(Tully, Wheat, et al., 2018)	10-20	PLM
Stordalen Mire thawing permafrost peatland(Martinez et al., 2019)	-10	PLS

#### 4.2 Calculation of the Arginine and Proline compositions in the ligase sequences' primary structures

The number and percentages of proline and arginine residues in the primary structures were examined on the Expasy server (<http://www.expasy.org/>) and PEPSTATS (EMBOSS)(<https://www.ebi.ac.uk/>). The proportion of arginine in the basic/positively charged amino acids group was also estimated on PEPSTATS(EMBOSS) (<https://www.ebi.ac.uk/>)server. The options for the two servers were set as default

#### 4.3 Analyzing the ligase sequences for globularity and disorder

The globularity (molecular packing) and loops (order and disorder regions) are part of the major factors that determine the protein thermostability (Barzegar et al., 2009). The globular domains in each ligase sequence were examined in the GlobPlot server. The Red Sea ligase sequences and other ligase sequences retrieved from NCBI nr database were used as a query on the GlobPlot server, with options set to default. The server can identify inter-domain segments consisting of linear motifs and other motifs that are not part of the ordered regions

of a recognized domain. GlobPlot assists in domain hunting efforts, predicting regions of enriched linear motifs, optimizing any constructs for expression, purification and crystallization and identifying the intrinsically disordered proteins (IDPs/IUPs) (Linding, Russell, et al., 2003). The Globplot makes use of the following algorithm:

$$Dis(a_i) = \sum_{i=1}^L P(a_i)$$

The  $P(a_i)$  is the propensity of the amino acid,  $L$  length of the amino acid,  $i$  is the natural logarithm and  $Dis(a_i)$  is the globularity detection.

The DisEMBL server (set to default) was used to explore the disorders in the loop/coil conformation of Red Sea ligase sequences and the other retrieved sequences NCBI (different bioproject) belonging to ligases from mesophilic environments (ML), ligases from thermophilic environments (TL), ligases from hyperthermophilic environments (HL) and ligases from psychrophilic environments (PL) groups. Each Red sea ligase sequence and the retrieved sequences were used as queries on the DisEMBL server which predicts the disordered regions within each protein sequence. Although a disordered state could have several definitions, this tool makes use of a term called ‘hot loops’ to signify coils with high-temperature factors. The hot loops are also referred to as the hotspots of the protein. Generally, protein disorder is normally found within loops. In DisEMBL, there is a DSSP program that differentiates the ‘loops’ from the ‘hot loops’ regions and calculates each amino residue on a two-state model, ordered or disordered (Linding, Jensen, et al., 2003).

#### 4.4 Analyzing the buried, exposed and functional proline and arginine residues

Red Sea ligase sequences and the other retrieved sequences (table 1) were used as queries in the ConSurf Server, with options set to default. ConSurf server makes use of a neural network algorithm to analyze the exposed, functionally exposed, buried, and structural buried amino acid residues in a protein sequence. The web server undergoes a six steps processes involving finding and aligning the sequences homologs, selecting the best evolutionary model, calculating the conservation scores using Bayesian model, searching the 3D structure for the protein sequence and projecting the conservation scores onto to the molecule (Ashkenazy et al., 2010)

#### 4.5 Proline and Arginine residues analysis in the loops and exposed regions of ligase sequences from different environment

The proline and arginine (PR) residues in the loop and hot loops of the Red Sea DNA ligase sequence were counted and compared to other DNA ligase sequences of their respective

genera that were recovered from different bio-projects on the NCBI. The loop and hot-loop analyses were made on the DisEMBL server as previously described. (Linding, Jensen, et al., 2003). The proline and arginine residues were also counted in the exposed and buried regions of secondary structures of each ligase sequence. The exposed-buried functional analysis was done on the ConSurf server as previously described.

## 5. Gene Synthesis, Cloning and Transformation

LigATL1 gene sequence from the Atlantis II LCL metagenomic datasets was modified by introducing BamHI and NdeI restriction sites to allow in-frame cloning into pET-16b (Novagen Wisconsin USA) with N-terminal 10x- His tag. This sequence was first codon-optimized for expression into *E.coli* using the Gene Script Codon Optimization tool before it was synthesized by Macrogen, Inc (South Korea). The gene was supplied in PUC 19 as a holding vector, and the lyophilized gene was stored at -4°C. Following the manufacturer's instruction, the plasmid DNA was centrifuged at 6,000 xg for 1 minute at 4°C, and 20 µl of sterilized water was added to dissolve the DNA. 1µl of plasmid DNA was taken from the stock and diluted ten times to a concentration of 20ng/µl. 3ul (60 ng/µl) of the plasmid DNA was transformed into competent *E.coli* DH5α and Top 10 cloning vectors using chemical heat-shock transformation following the instructions of Chang *et al.*, 2017. The transformants were plated on 100µg/ml ampicillin LB agar plates.

The Plasmid DNA was extracted from a freshly prepared overnight culture of the positively transformed clone of *E coli* DH5 α using QIAprep Spin Miniprep Kit (Qiagen) according to the manufacturer's instruction. The plasmid DNA was digested by restriction enzymes BamHI and NdeI following the NEB standard protocol, after which the size of the insert (ligase gene) was confirmed by running it on 1% agarose gel. In addition, the correct orientation and the size of the insert were also confirmed by digesting with KpnI and NdeI in order to cut within the insert (ligase gene). The ligase gene was gel purified using Zymoclean™ Gel DNA Recovery Kit (Zymo Research, Irvine, CA, USA) following the manufacturer's instructions. The concentration of the ligase gene was determined using NanoDrop™ 2000/2000c Spectrophotometers. Using the NEB standard protocol, the ligase gene was cloned into pET-16b (Novagen). The 15µl reaction mixture contained 3:1 gene to vector molar ratio and 2U of T4 ligase, and the reaction was incubated overnight. The ligation was confirmed by running on 1% agarose gel. The recombinant plasmid was transformed into chemically competent *E coli* BL21(DE3) and *E coli* BL21 PLYS for protein expression according to similar instructions from Chang *et al.*, 2017

## 6. Expression of Lig-ATL1 in *E coli* BL21(DE3) and *E coli* BL21 PLYS

LigATL1 was induced in *E.coli* BL21(DE3) expression host firstly, by preparing an overnight culture of the transformed cells grown in LB medium ( Lysogeny broth, containing 10g of tryptone, 5g of Yeast extract and 10g of NaCl) (Gibco<sup>R</sup>, lennox, New York) which also contained 100µg/ml ampicillin and placing it on a shaking incubator ( innova<sup>R</sup> 43, New Brunswick) at 200rpm and 37°C. Afterward, 20 ml of this culture was inoculated in 1 L of LB

plus 100µg/ml of ampicillin that was distributed in 250ml fractions into four baffled 1L flasks for proper aeration. The cultures were grown inside a shaking incubator at 37°C, 200rpm until the OD<sub>600</sub> (Optical density) reached approximately 0.6, after which 1ml of aliquot was removed before inducing with 0.1mM of Isopropylβ -D-1 thiogalactopyranoside (IPTG). The cultures were then re-incubated at 37°C, 200rpm for 5hrs induction. The pellet was harvested by centrifuging at 8,000rpm, 4°C for 15 mins and stored at -80°C. The 50 ml of the cell pellet was lysed on ice using SONIFER<sup>®</sup> 150, Branson for 3 mins, but 250ml of the cell pellet was used under the large-scale expression. The cell lysate and the pellet were analyzed for solubility of protein expression by running on 12% Sodium dodecyl sulfate-polyacrylamide gel electrophoresis according to Laemmli's standard protocol (Gallagher, 2006). The same set of procedures were done with the expression of LigATL1 in *E.coli* BL21 PLYS.

#### **7. Partial Purification of His-tagged Lig-ATL1 on Ni<sup>2+</sup> column**

The cell pellets were incubated on ice to thaw and then resuspended in binding buffer (20mM, NaH<sub>2</sub> PO<sub>4</sub>, 0.3M NaCl, pH 8.0, 10% glycerol, 0.2% Triton X). The cells were sonicated on ice using SONIFER<sup>®</sup> 150, Branson for a total of 9 mins (250ml cell pellets, burst for every 30 secs and then separated by 30 secs pause). The supernatant of the cells was separated from the cell debris after centrifugation at 4°C, 11,000rpm for 40mins. The Lig-ATLI protein was partially purified using Ni<sup>2+</sup> affinity chromatography on a 1ml- Ni-NTA agarose resin (GE Healthcare column). The purification protocol was followed according to the manufacturer's instructions with the exception of no imidazole in the binding buffer and 30mM of imidazole concentrations in the Elution buffer (20mM, NaH<sub>2</sub> PO<sub>4</sub>, pH 8.0, 30mM imidazole, 0.3M NaCl, 50% glycerol). Several purification trials were done previously to reach the best imidazole concentrations in the binding and elution buffers. The purification of the protein was visualized on 12% SDS PAGE gel at various intervals. The SDS PAGE preparation, procedures, and buffers were according to Laemmli's standard protocol (Gallagher, 2006).

## Chapter 3: Results and Discussions

### 1. Contig Assembly

The DNA isolated from the several Red Sea brine pools (Atlantis II Deep brine pool (LCL), Discovery Deep brine pool and Kebrit Deep Upper and Lower interface layer) was pyrosequenced using Roche-454. The assembling of reads from each Red Sea brine pool metagenomes resulted in contigs, and this result is summarized in table 1.

**Table 2: Assembling of reads of Red Sea metagenomes to contigs using MEGAHIT**

Red Sea Brine Pools	No of Reads	No of Contigs(bp)	Avg. contig Length(bp)	Contig length Range(bp)
Atlantis II Deep (LCL)	2,162,162.25	107,740	609.24 ± 426.42	46605
Kebrit Deep Lower Interface	1,510,272.25	91,348	595.26 ± 366.02	10550
Kebrit Deep Upper Interface	1, 562, 521.25	230,017	597.98 ± 426.42	16986
Discovery Deep	913,497.25	187,457	591.03 ± 314.93	9197

### 2. DNA Ligase Screening from the Red Sea brine pools Metagenomic datasets

Open reading frames (ORFs) were predicted using the Metagene Annotator (Noguchi et al., 2008) using the final contigs in order to perform HMM search with the HMM model built from several raw HMMs of the 7 DNA ligase domains (N terminus (PFO4675), C terminus (PFO4679), ATP dependent DNA ligase (PFO1068), Adenylation (PFO1653), Oligonucleotide Binding, OB (PF14743), Zinc fingerlike Binding, ZBD (PFO3119) and BRCA 1 C terminal, BRCT (PF0533)) retrieved from the Pfam database 32.0. 18 ORFs for putative DNA ligase were selected based on certain criteria which include significant low e-value and length of the ORFs. The ORFs were annotated using nrBlast P, and the results are presented in the table 2. Furthermore, Two of the ORFs (ATLORF1, ligATL1) and KDUORF4, ligKDU4 were further selected as potential DNA ligases for synthesis based on the presence of all the essential DNA ligase motifs and their identities to already existing genes. LigATL1 was first synthesized for cloning and expression. LigATL1 and ligKDU4 annotations



were confirmed on the Conserved Domain Database (CDD). Lig ATLI, Lig KDU4 and many other Red Sea brine pools ORFs for putative DNA ligase were found in contig K141 assembled through MEGAHIT(D. Li et al., 2016).

**Table 3: Putative ORFs for DNA ligase sequences selected from the predicted ORFs of Red Sea brine pools final Contigs**

**3a) Putative DNA ligases (ATP type) from Atlantis II brine pool assembled final contigs**

ORFS	Best Hits (ATP type),	E-values	Identities	Query Covers	Super Family
ORF 1, 298	<i>Erysipelotrichaceae bacterium</i>	6e-52	42.1%	92%	CDC9
ORF2, 244	<i>Methanomassiliicoccales archaeon</i>	2e-17	34.2%	72%	CDC9
ORF3 541	<i>Phyllobacterium myrsinacearum</i>	0.0	95.19%	100%	Lig bact
ORF4 637	<i>Rhizobiales bacterium,</i>	0.0	81.04%	99%	LigD
ORF5, 214	<i>Rhizobium phaseoli,</i>	1e-81	67.62%	95%	LigD
ORF6, 163	<i>Cupriavidus metallidurans</i>	2e-60	62.73%	98%	CDC9
ORF7 164	<i>Thermoprotei archaeon</i>	9e-37	47.3%	88%	CDC9

### 3b) Putative DNA ligases (NAD<sup>+</sup> type) from Atlantis II brine pool assembled final contigs

ORFs	Best Hits (NAD type),	E-values	Identities	Query Covers	Super-Family
ORF 8, 714	<i>Phyllobacterium myrsinacearum</i>	0.0	98.88%	99%	Lig A
ORF 9, 459	<i>Moraxella osloensis</i>	0.0	98.86%	95%	Lig A
ORF 10, 444	<i>Candidatus Marinimicrobia bacterium,</i>	0.0	78.20%	94%	Lig A
ORF 11, 208	<i>Moraxellaceae bacterium,</i>	3e-138	97.6%	99%	Nil

### 3c) Putative DNA ligases from Discovery Deep brine pool (DDP) and Kebrit Deep Upper (KDU) and Kebrit Deep Lower Surfaces assembled final contigs

ORFs	Best Hits	E-values	Identities	Q. covers	Super Family
DDP 1 305	ATP-dependent DNA ligase [ <i>Prochlorococcus marinus</i> ]	0.0	96.72%	100%	CDC9
DDP 2 235	hypothetical protein DRO61_10175 [ <i>Candidatus Bathyarchaeota archaeon</i> ]	2e-142,	87.11%	95%	CDC9
KDL 397	NAD-dependent DNA ligase LigA [ <i>Candidatus Portnoybacteria bacterium</i> ]	9e-168	58.33%	99%	LigA
KDU 1 338	NAD-dependent DNA ligase LigA [ <i>Candidatus Marinimicrobia bacterium</i> ]	0.0	89.94%	100%	LigA
KDU 2 567	DNA ligase (NAD(+)) LigA [ <i>Acidimicrobiaceae bacterium</i> ]	0.0	98.94%	100%	Lig A
KDU 3 319	DNA ligase (NAD(+)) LigA [ <i>Planctomycetes bacterium GWF2_40_8</i> ]	0.0	88.82%	98%	Lig A
KDU 4 631	DNA ligase (NAD(+)) LigA [ <i>Candidatus Marinimicrobia bacterium</i> ]	0.0,	63.55%	100%	LigA

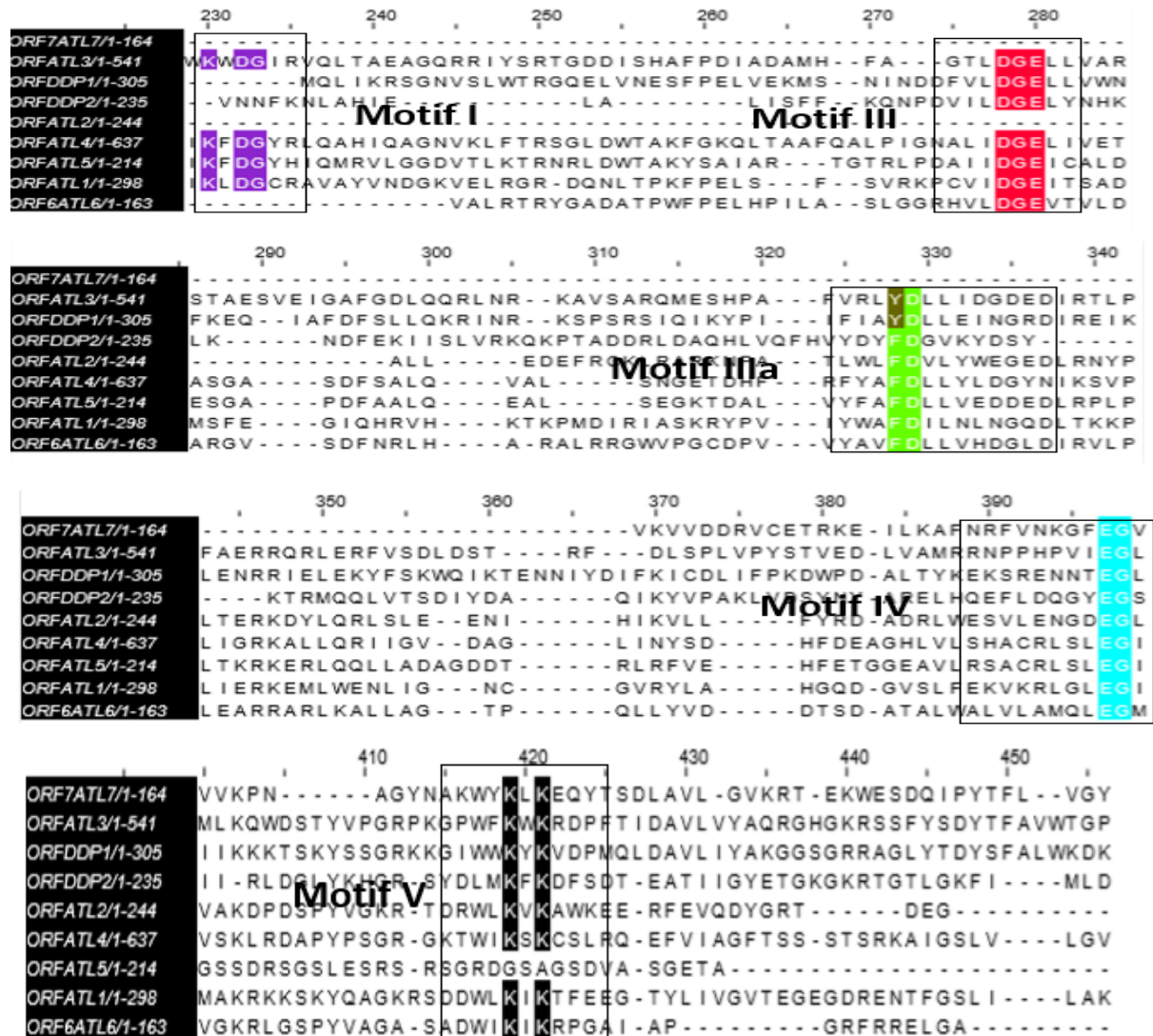
### 3d) The ORFs selected for Synthesis

2.1

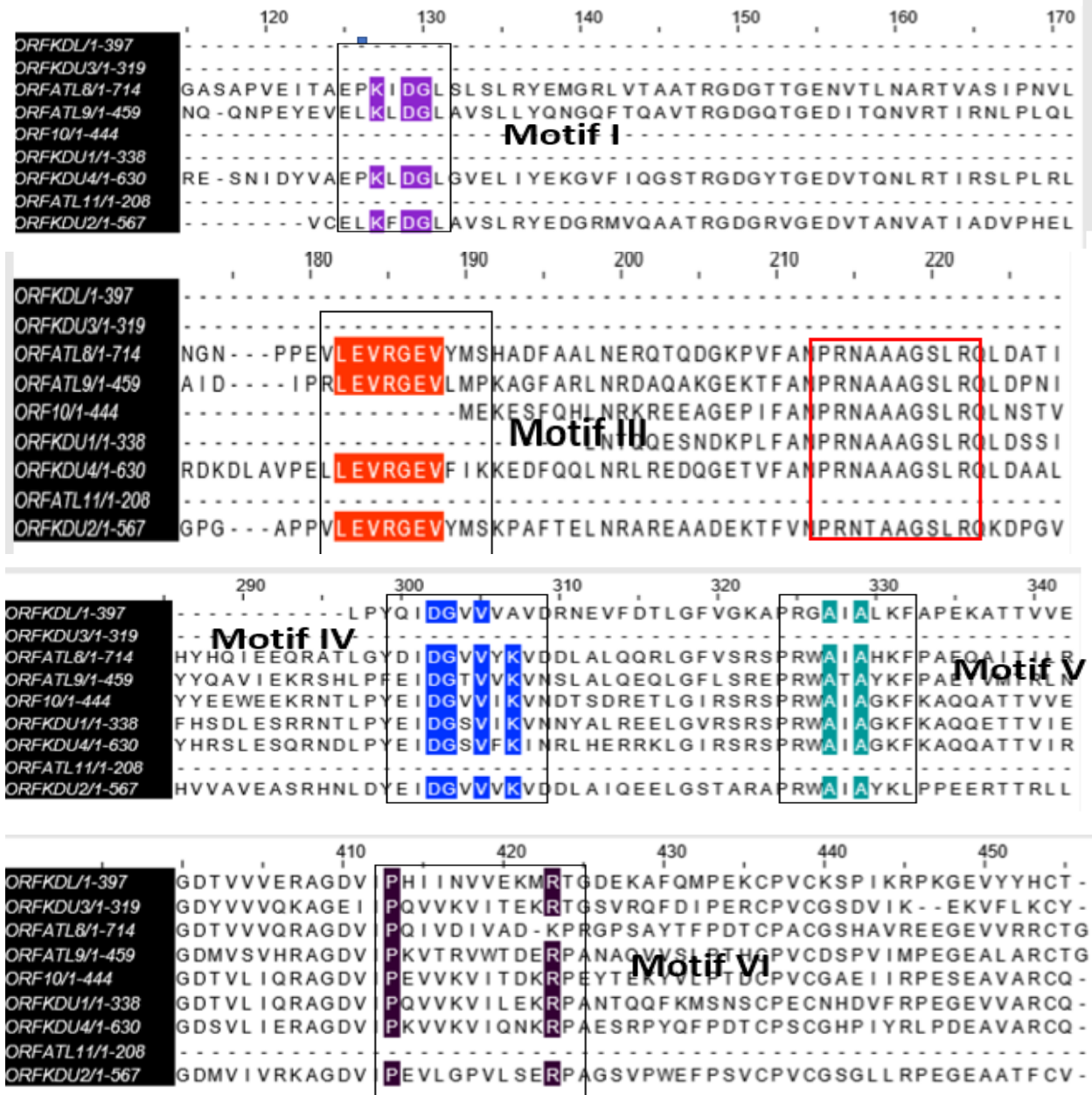
Database	ATLII ORF1	KDUORF4
<b>nrBlast P</b>		
E values	6e-52	0.0
Description	ATP type from <i>Erysipelotrichaceae bacterium</i>	NAD+ type from <i>Candidatus Marinimicrobia bacterium</i>
% identities	42%	62.55%
Hit coverage	92%	100.0%
<b>CDD search</b>		
E-value	6.52e-62	0e+00
Interval	1-291	1-630
Accession	COG1793 (CDC9)	PRK07956(LigA)
Description	ATP-dependent DNA ligase [Replication, recombination and repair]	NAD+ DNA Ligase(LigA)

### Multiple Sequence Alignment (MSA) of all the Selected DNA Ligase ORFs from Red Sea Metagenomic Datasets

Red Sea brine pools DNA ligase ORFs were grouped into NAD<sup>+</sup> and ATP types based on Blast P annotation and were aligned separately. A series of conserved motifs for DNA ligases were detected in the ORFs with some of them showing all the conserved essential motifs(Fig.8 & Fig.9). Motifs 1-V of DNA ligase are found in the core domains of the ligase enzyme, and they are important for nucleotide binding. Also, several biochemical and genetic studies have also suggested that the conserved residues are critical for the DNA end-joining process (Wilkinson et al., 2001). The Motif I, KXDG (purple) contains the active site lysine residue which is involved in covalent adduct formation with AMP. Motif III has a glutamate residue(red) which forms a hydrogen bond with the ATP ribose unit and Motif IIIa contains essential tyrosine that stabilizes the adenine ring. Motif V has an essential lysine that brings the  $\alpha$ -phosphate group in contact(Doherty & Suh, 2000). Red Sea brine pools ligase ORFs belonging to NAD<sup>+</sup> dependent DNA ligase and ATP-dependent DNA ligase showed minimal identity, but they have some homology in the conserved Motifs 1-V (figure 8, figure 9). This indicates that all DNA ligases follow the same pattern of reaction mechanism(Wilkinson et al., 2001). The ORFs of putative NAD<sup>+</sup> Dependent DNA ligases have the most conserved residues (PRNAAAGSLR) of the NAD<sup>+</sup> Dependent DNA ligase family (figure 8 & figure 10). This is confirmed further by aligning LigATL1 and LigKDU4 to bacterial DNA ligases (figure 10).

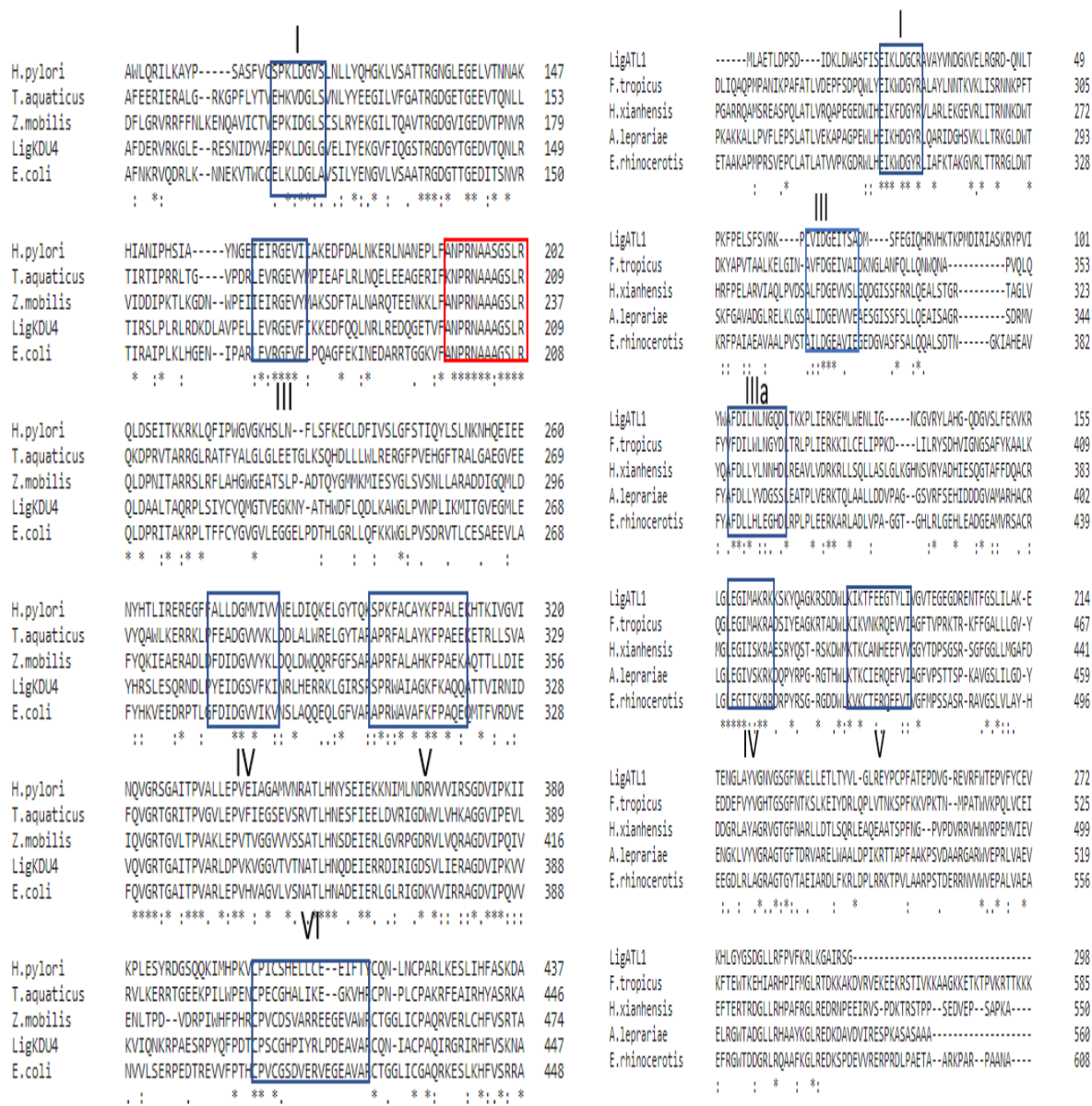


**Figure (8): Essential Motifs in ORFs coding for Putative ATP Dependent DNA Ligases from the Red Sea Metagenomes:** The motifs (I-V) are put in a rectangle with their conserved residues colored. The active site motif of DNA ligases, KXDG is highlighted in purple, ORFATL 2, ORFATL6 and ORFDDP1 lack the KXDG motif but have all other motifs present. CLUSTAL omega was used to generate the alignment.



**Figure (9): Multiple Sequence Alignment showing the essential Motifs in ORFs coding for Putative NAD<sup>+</sup> Dependent DNA Ligases from the Red Sea Metagenomes:** The motifs (1-VI) are put in a rectangle with their conserved residues colored. The active site motif of DNA ligases, KXDG is highlighted in purple, ORFs KDL, KDU3, ATL1 and ATL 10 lack motif 1 and III but contain all other motifs. The most conserved motifs in all NAD<sup>+</sup> Dependent DNA ligases including that from the Red Sea is boxed in the red rectangle. CLUSTA omega was used to generate the multiple sequence alignment





**Figure (10): Alignment of LigKDU4 and LigATL1 with the conserved sequences among Bacterial DNA ligases ( NAD<sup>+</sup> type to the right and ATP type to the left): five conserved motifs (1-V) identified in both NAD<sup>+</sup> and ATP DNA ligases, motif VI which is also present in mRNA capping enzymes is seen in NAD<sup>+</sup> DNA ligase type. NAD<sup>+</sup> DNA ligase most conserved residues (PRNAAAGSLR) is boxed in the red rectangle, and these residues are never present in the ATP DNA ligase type. The NAD<sup>+</sup> DNA ligases (Lig A) sequences selected are from *H. pylori*; *Helicobacter pylori* 2017, *T.aquaticus*; *Thermus aquaticus*, *E.coli*; *Escherichia coli*, The ATP Dependent DNA ligase (Lig D) sequences selected are form, *H.xianhensis*; *Halomonas***

*xianhensis*, *F.tropicus*; *Flavisolibacter tropicus*, *A. leprariae* *Aureimonas leprariae*, *E. rhinocerotis*; *Enterovirga rhinocerotis*. CLUSTAL omega was used to generate the alignment

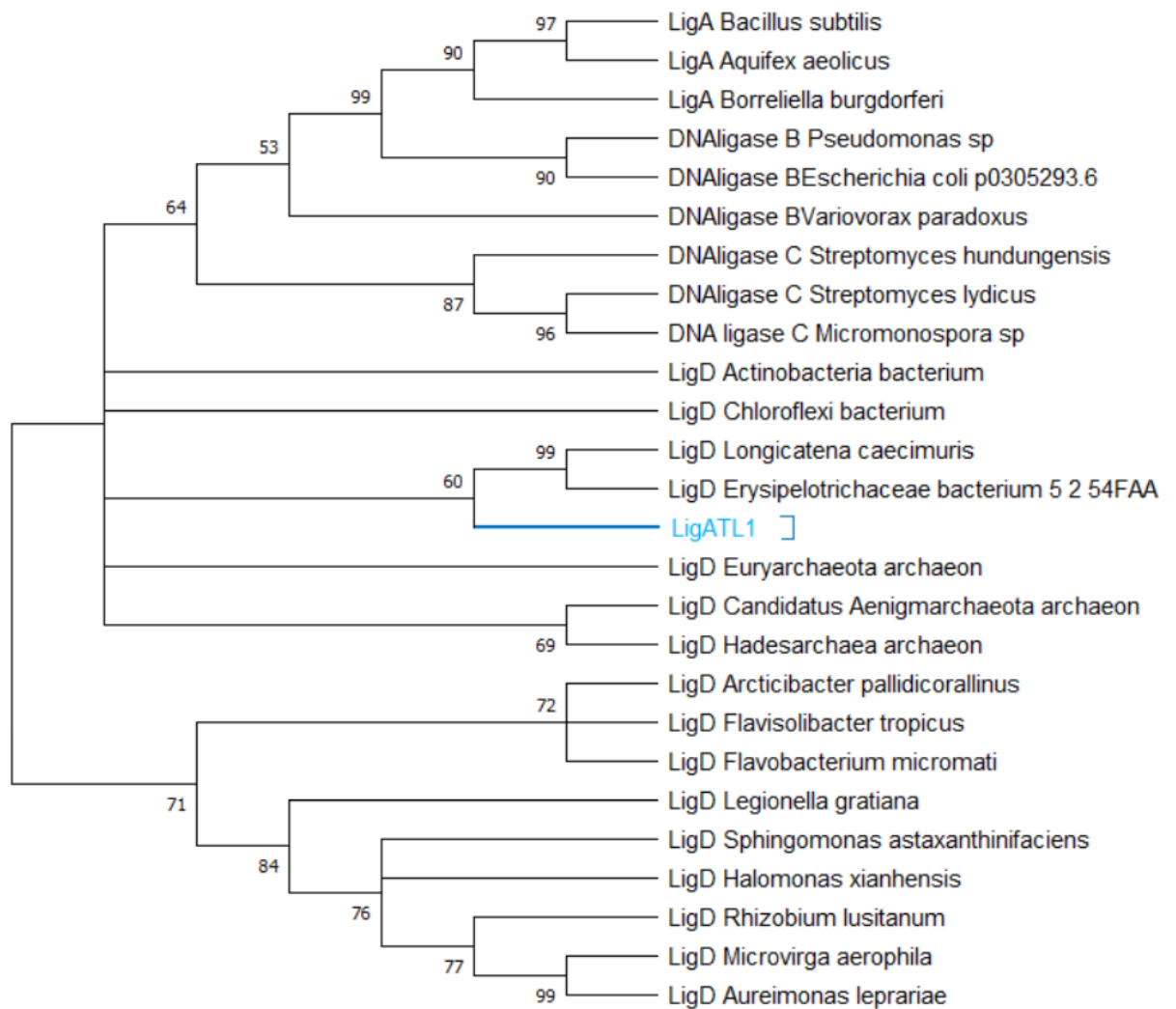
## 2.2 Phylogenetic Analysis of LigATL1 and LigKDU4

The evolutionary study suggests that bacteria have both the ATP and NAD<sup>+</sup> Dependent DNA ligases, while ATP Dependent DNA ligases are found exclusively in Archaea, Eukaryotes and Virus(Pergolizzi et al., 2016; Wilkinson et al., 2001). The 68 Ligase sequences that were acquired from the NCBI database contained different types of ligases; the eukaryotic and virus ATP dependent DNA ligases (DNA ligase I, III and IV), Bacteria and Archaea ATP dependent DNA ligases (lig B, C & D) and Bacteria NAD<sup>+</sup> dependent DNA ligase (Lig A)(fig 11a). These ligases were aligned, and Phylogenetic analysis was performed. The phylogenetic tree was reconstructed with LigATL1, LigD sequences and few LigB and Lig D sequences due to low bootstrapping value for LigATL1 in the first phylogenetic tree(fig11b).

The tree topology showed that NAD<sup>+</sup> Dependent DNA ligase clustered into one clade while the ATP dependent DNA ligases varied and expanded into many clades. LigATL1 was grouped amongst the LigD family which are ATP Dependent DNA ligases, while LigKDU4 was present amongst the LigA family which are NAD<sup>+</sup> Dependent DNA ligases. The eukaryotic ATP dependent DNA ligase (DNA ligase I, III and IV) appeared to come from the same root or descended from a common ancestor. Bacterial ATP-dependent DNA ligases (Lig B, C & D) have diverse clades possibly because they have evolved from multiple gene transfer events that have taken place either within bacteria species, bacteria and archaea or integration of viral DNA into bacterial genomes(Sharma et al., 2009). The complex evolution of bacterial ATP-dependent DNA ligase is debatable because scientists argued whether ATP dependent or NAD<sup>+</sup> DNA ligase was the first to evolve(Cole et al., 2001; Pergolizzi et al., 2016).







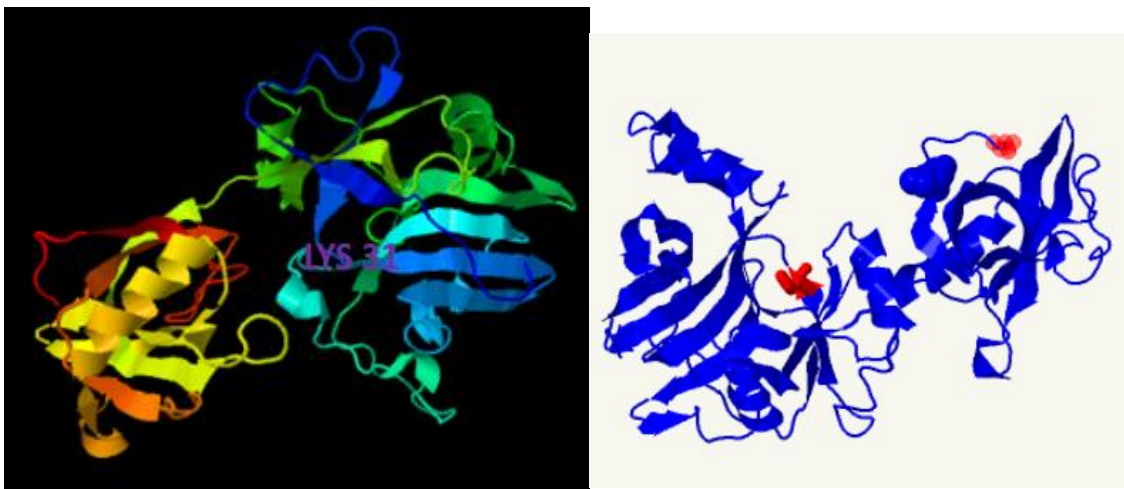
**Figure 11b: Reconstructed phylogenetic tree for LigATL sequence.**

Ligase A, B, C & D families were used in the tree construction and the bootstrapping test was done in 1000 replicates. Branch numbers indicate the bootstrap values. LigATL is colored and underlined in blue. Megahit was also used in reconstructing the phylogenetic tree. MEGA HIT was used in generating the tree and was condensed to only include bootstrap values > 50%

### 3. In silico Characterizations of LigATL1 and LigKDU4

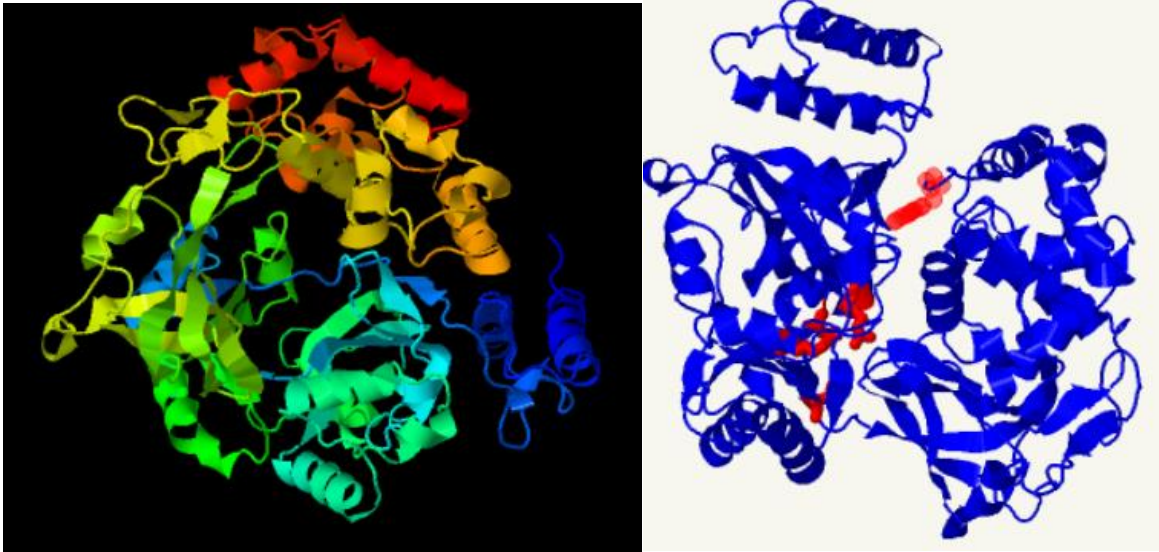
#### 3.1 3D Modelling and Predicting the Superimposition of LigATL1 (ATP type) and LigKDU4 (NAD<sup>+</sup> type)

LigATL1 and LigKDU4 were structurally predicted using the PHYRE2 Protein Fold Recognition Server. LigATL1 was modeled at 100% confidence using a bound adenylated human DNA ligase as a template and the coverage was over 90%. LigKDU4 was also modeled at 100% confidence with a bound adenylated NAD<sup>+</sup> *E.coli* DNA ligase and the coverage was also over 90%. The 3D structure revealed that LigATL1 has the Adenylation domain and the OB domain of LigD while LigKDU4 contains the four full domains (Adenylation, OB-fold, Zn finger and HhH) of Lig A. The N terminal domain of ligATL1 is from 1-180 residues and that contains the active site and the adenylation domain while the C terminal domain extends from 181 to 298 residues and contains the OB fold domain. The LigKDU4 has its N terminal domain from 1-440 residues, and it contains the adenylation domain and OB domain. The C terminal domain extends from 441-631 residues and has the Zn finger-like and the HhH domains. Thermostable DNA ligases from *S.solfataricus* and *t2. filliformis* were discovered in the best hits of LigATL1 and LigKDU4, respectively, and both superimposed well with a confidence of 100% and Template modeling <sup>TM</sup> score of 1.00(fig.12& fig.13). *Sulfolobus solfataricus* and *Thermus filliformis* are thermophiles and the optimum temperature of thermostable DNA ligase in *t2. filliformis* is estimated to be 70° C(Jeon et al., 2004) The superpositions on the PHYRE 2 server were generated by the largest subset of atoms that are superimposable within some threshold (3.5 Å) and which were then adjusted based on the size of the proteins being aligned. The PHYRE2 server predicts models with high similarity having Template Modeling score (TM-score) > 0.7, average similarity has same fold > 0.5, and no similarity have different folds < 0.5.

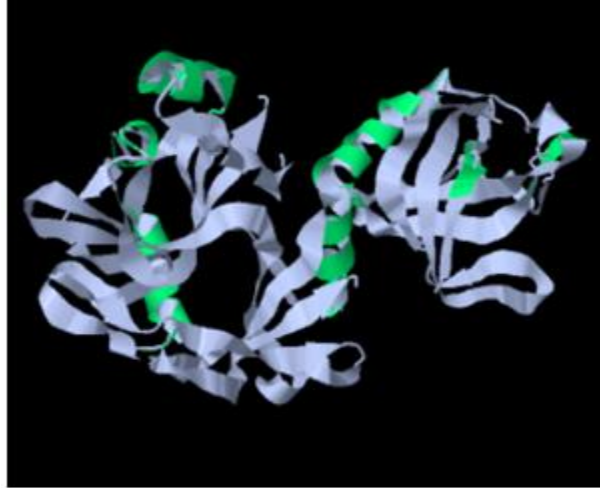


**Figure (12): Modelling of LigATL1 using a crystal structure of human DNA ligase bound to 5'-adenylated, nicked2 DNA as a template (confidence 100%, identity 28%).** The 3D structure of LigATL1 has the N terminal/ Adenylation domain which consists of subdomain a

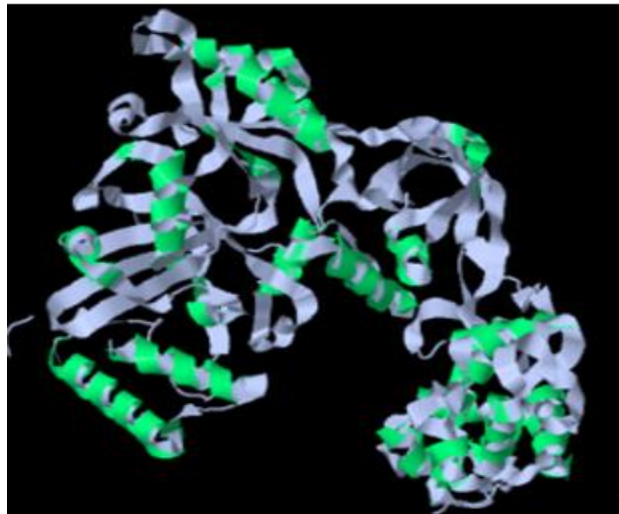
(blue) and subdomain b (green/cyan). The C terminal domain contains the OB-fold domain (yellow) and zinc finger-like (red) domain. The catalytic site Lys 31, according to the Catalytic Atlas Structure, CAS site lies in the N terminal domain is indicated, and it is shown in red in the blueprint (right image). The 3D structure prediction was generated using PHYRE2 Protein Fold Recognition Server (Kelley & Sternberg, 2015). The image was produced by JSmol.



**Figure (13): Modelling of ORFKDU4 with NAD+ *E.coli* DNA ligase bound2 to nicked DNA-adenylate, (confidence 100%, identity 50%).** The 3D structure of LigKDU4 has an Adenylation domain that is subdivided into domain 1a ( light blue) and domain 1b ( green/cyan), domain 2( yellow); OB-fold, sub-domain 3a(red); Zinc finger domain, and sub-domain 3b(orange); HhH. LigKDU4 has 4 catalytic sites (K114, D116, R201, K310) according to the Catalytic Atlas Site, CAS site, and they are shown in red in the blueprint (right image). The 3D structure prediction was generated using PHYRE2 Protein Fold Recognition Server (Kelley & Sternberg, 2015). The image was produced by JSmol.



**Figure (14): Superimposition of LigATL1 with thermostable ATP-dependent DNA ligase from *S. solfataricus*, (identity 31%).** The ligase is superimposed at a confidence level of 100% with the highest template modeling TM score (1.0). 304 residues are paired between both ligases. The gray represents the template (*S.solfataricus*) and the green is LigATLI. The superimposition was predicted on the PHYRE2 server and the image was generated by JSmol.



**Figure (15): Superimposition of LigKDU4 (Kebrit Deep) with thermostable *t2.filliformis* DNA Ligase, superimposed at confidence at 100% with 48% identity.** The ligase showed the highest similarity possible (TM score =1) to the template. 586 residues are paired between both ligases. The gray represents the template (*t2.filliformis* DNA ligase) and the green represents LigKDU4. The superimposition was predicted on PHYRE2 server and the image was generated by JS mol



#### 4. Predicting the Molecular weight and Isoelectric point (pI) of LigATL1 and ligKDU4

The theoretical Isoelectric point (pI) and the Molecular weight were initially calculated in the Expasy server (<https://web.expasy.org/protparam/>) and while troubleshooting the solubility expression of LigATL1( 2 units pH difference between protein and lysis buffer) , the pI was re-calculated on another server(isoelectric.org) . The pIs of 7. 59 and 5.55 and molecular weights of 33.76kDa and 71.89kDa were estimated for LigATL1 and LigKDU4, respectively, on the Expasy server. The pI and molecular weight of LigATL1 were 6.98 and 33.76kDa, while that of ligKDU4 were 5.39 and 71.23kDa on the isoelectric.org(table 4). The difference in estimated theoretical pI of LigATL1 between both servers may be from changes in the algorithm used. Generally, the idea of pI is to count the charged amino acids and iteratively calculate the Henderson-Hasselbach equation (equation to estimate pH of a buffer solution) on them. Expasy server did not consider Histidine as positively charged residue and so neglected it in the pI calculation, while in isoelectric.org, histidine is part of counted charged residues and thus, calculating a lower pI.

**Table 4: Isoelectric points and Molecular weights calculations of LigATL1 and LigKDU4**

**sequence**

<b>Calculations on the servers</b>	<b>LigATL1</b>	<b>LigKDU4</b>
Isoelectric point, pI (Expasy server)	7.59	5.55
Molecular weight(kDa) (Expasy server)	33.76	71.89
Isoelectric point, pI (isoelectric.org server)	6.98	5.39
Molecular weight(kDa) (isoelectric.org server)	33.76	71.23

#### 5. Comparative thermostability analysis of Red Sea ligase sequences to ligase sequences of different environments: stability roles of proline and Arginine.

Orthologous proteins can evolved based on differences in thermal stability which may be from the natural selection that favor some amino acid usages which then optimize protein function to better adapt to a particular thermal environment (Chao et al., 2020; Fields, 2001; Sælensminde et al., 2009).The temperature has a strong impact on the stability and flexibility of protein, and this makes the protein susceptible to unfolding and stability reduction when exposed to unfavorable environmental temperature (Goldenzweig & Fleishman, 2018; Somero et al., 2017).Biases in amino acid usages are often associated with disordered regions of protein which promote adaptation to extreme environment(Dunker et al., 2001). Disordered regions are enriched with some amino acids such as Pro, Arg, Ala, Lys, Gln, Glu, Ser and Gly and depleted in order promoting amino acids like Trp, Cys, Phe, Ile, Val and Leu (Dunker et al., 2001; Uversky et al., 2000). Studies involving protein engineering have shown that proteins from thermophilic and hyperthermophilic environment have higher frequency of proline and arginine residues at certain loop regions(Barzegar et al., 2009; Georlette et al., 2000; Igarashi et al., 1999; Matsutani et al., 2011; Watanabe et al., 1991)

In this comparative thermostability study, ligase sequences were grouped under organisms of closest taxonomic groups that are likely to be found in one or more extreme environments to imitate orthologs sequences based on thermal environment. For instance, in the *Candidatus Marinimicrobia bacterium* group, ligase sequences were retrieved from the metagenomic assemblies of Guayamas base sediment and cold, oxic seafloor aquifer that were predicted to have *Candidatus Marinimicrobia* species in their metagenomes. Nine Red Sea putative ligase sequences were also included in this study, five from Atlantis II LCL layer(ORF8(TLA1), ORF4(TLA2), ORF3(TLA3), ORF9(TLA4) and ORF10(TLA5) and four from Kebrit Deep interfaces( KDU4(MLK),KDU2(MLK2), KDL(MLK3) and KDU1(MLK4). Others putative Red Sea ligase sequences were exempted in the comparative thermostability study either because they have shorter lengths residues or did not fit a group (no ligase sequences from other extreme environments to compare them with). Proline and arginine residues at the disordered and exposed regions were examined in total of 22 ligase sequences placed under 5 different closest taxonomic groups (table 5).

### **5.1 Arginine and Proline Compositions in the Primary Structures of the ligase sequences**

The number and percentages of proline and arginine residues in each ligase sequence primary structure (amino acid residues) are summarized in table 5. Also, the ratio of arginine residues to the total basic/positively charged amino acid residues are reported in the same table. The five taxonomic category grouped based on temperature driven orthology include: *Candidatus Marinimicrobia bacterium*(CBA&b group), *Acidimicrobiaceae bacterium* group (AB group), *Moraxallaceae bacterium* group (MB group), *Phyllobacterium myrsinacearum* group (PM group), *Rhizobiales bacterium* group (RB group)



**Table 5: Pro and Arg Compositions of ligase sequences used in the Comparative study**

DNA ligase.	Env.temp(°C).	P(%)	R(%)	R/ R+K+H	Closest taxonomic group
<b>Candidatus Marinimicrobia bacterium group Sp.1 (CBa group)</b>					
<b>HLG</b> (670)	>120	36(5.4)	47(7.0)	0.45	<i>Candidatus Marinimicrobia bacterium</i>
<b>MLK</b> (630)	23.3	29 (4.6)	57(9.0)	0.58	<i>Candidatus Marinimicrobia bacterium</i>
<b>PLC</b> (668)	4	27 (4.0)	50(7.5)	0.54	<i>Candidatus Marinimicrobia bacterium</i>
<b>Acidimicrobiaceae bacterium group (AB group)</b>					
<b>HLH</b> (664)	95	41(6.2)	72(10.8)	0.75	<i>Acidimicrobium ferrooxidans</i>
<b>TLY</b> (681)	43	34(5.0)	48(7.0)	0.54	<i>Ferrithrix thermotolerans</i>
<b>MLK2</b> (567)	23.3	32(5.7)	41(7.3)	0.67	<i>Acidimicrobiaceae bacterium</i>
<i>E.coli</i> (671)	37	30(4.5)	46(6.9)	0.49	<i>Escherichia coli</i>
<b>PLB</b> (684)	1-10	41(6.0)	41(6.0)	0.51	<i>Acidimicrobium sp.</i>
<b>PLB2</b> (491)	1-10	26(5.3)	31(6.3)	0.42	<i>Acidimicrobiaceae bacterium</i>
<b>Moraxallaceae bacterium group (MB group)</b>					
<b>TLA4</b> (459)	68	25(5.4)	29(6.3)	0.52	<i>Moraxella osloensis</i>
<b>MLT2</b> (455)	20-25	17(3.7)	24(5.3)	0.42	<i>Moraxellaceae bacterium</i>
<b>PLC2</b> (427)	4	16(3.7)	21(4.9)	0.36	<i>Moracellaceae bacterium</i>
<b>Phyllobacterium myrsinacearum group (PM group)</b>					
<b>TLA1</b> (714)	68	32(4.5)	52(7.3)	0.547	<i>Phyllobacterium myrsinacearum</i>
<b>TLA3</b> (541)	68	33(6.1)	51(9.4)	0.637	<i>Phyllobacterium myrsinacearum</i>
<b>MLT</b> (714)	20-25	32(4.5)	51(7.1)	0.542	<i>Phyllobacterium myrsinacearum</i>
<b>Rhizobiales bacterium group (RB group)</b>					
<b>TLA2</b> (637)	68	57(5.8)	41(6.4)	0.41	<i>Rhizobales bacterium</i>
<i>E.coli</i> (671)	37	30(4.5)	46(6.9)	0.49	<i>Escherichia coli</i>
<b>PLM</b> (710)	10-20	26(3.6)	29(4.0)	0.24	<i>Rhizobales bacterium</i>
<b>Candidatus marinimicrobia bacterium Sp.2 (CBb group)</b>					
<b>HLG2</b> (398)	>120	19(4.8)	30(7.5)	0.462	<i>Candidatus Poribacteria bacterium</i>
<b>TLA5</b> (444)	68	19(4.3)	34(7.7)	0.576	<i>Candidatus Marinimicrobia bac.</i>
<b>MLK3</b> (397)	23.3	17(4.3)	19(4.8)	0.279	<i>Candidatus Portnoybacteria bac.</i>
<b>MLK4</b> (338)	23.3	14(4.1)	22(6.5)	0.45	<i>Candidatus Marinimicrobia bac.</i>
<b>PLS</b> (395)	-10	16(4.1)	26(6.6)	0.448	<i>Candidatus crysoericum odellii</i>

\*P is the number of proline residues, R is the number of arginine residues, K and H are lysine and histidine, respectively. **HLG(1&2)**: ligase sequences from the Hyperthermophilic hydrothermal Guayamas base sediment metagenomic assembly(Seitz et al., 2019), **HLH**: ligase sequence from the Hyperthermophilic Hot spring Iceland(Clum et al., 2009), **TLA**: Ligase sequences from Atlantis II, LCL metagenomic assembly **MLK**: ligase sequence from the Mesophilic Kebrit Deep upper surface metagenomic assembly, **MLT**: Ligase sequence from mesophilic terrestrial habitat(forest)(Tatusova et al., 2014), **PLC**: ligase sequence from the Psychrophilic Cold, oxic subseafloor metagenomic assembly(Tully, Wheat, et al., 2018), **TLY**: ligase sequence from moderately thermophilic mineral samples from the Yellowstone National Park United States(Tatusova et al., 2014), **PLB**: Ligase sequence from Psychrophilic Russia lake Baikal sub

ice water metagenome(Cabello-Yeves et al., 2018), **PLB2**: Ligase sequence from Psychrophilic Japan lake Biwa freshwater metagenome(Hiraoka et al., 2019), **PLM**: ligase sequence from the Mediterranean sea, Marine Plankton Metagenomic assembly(Tully, Graham, et al., 2018). **PLS**: Ligase sequence from metagenome-assembled from Stordalen Mire thawing permafrost peatland(Martinez et al., 2019). Blue code for ligase sequences from hyperthermophilic environment, red code for thermophilic, orange for mesophilic and black for psychrophilic.

Increase in the number of proline residues have been associated with many thermophilic enzymes, and this allows for more stable conformation at high temperature (Barzegar et al., 2009; Igarashi et al., 1999; Watanabe et al., 1991). These proline residues provide more rigidity and reduce flexibility in the thermostable protein environment(Georlette et al., 2000). Proline is a unique amino acid because of its pyrrolidine ring that disrupts protein tertiary conformation and reduces the conformational entropy, thereby enhancing the thermostability of protein(Barzegar et al., 2009). Many ligase sequences from thermophilic and hyper-thermophilic environments such as HLG 1 & 2, HLH TLA5, TLA 4 and TLA2 have higher percentages of proline in their primary structures(amino acid residues) compared to similar ligase sequences of closed taxonomic groups from other environments(table 5).

Arginine contributes to the thermostability of protein because of its three nitrogen atoms from its guanidinium group that forms more hydrogen bonds and salt bridges than many other amino acid residues (Barzegar et al., 2009; Georlette et al., 2000). Comparative genomic study of thermophilic and mesophilic acetic acid bacteria, AAB strain(*Acetobacter tropicalis*) has shown that massive amino acid substitutions of Lys to Arg residues occurred, and the number of Arg-based salt bridges increased in thermophilic strain(Matsutani et al., 2011). Codon usage that preferentially selects arginine in protein translation has also been noticed in many thermophiles and hyperthermophiles(Van der Linden & de Farias, 2006). An increase in Arg residues in the primary structure of some mesophilic enzymes compared to thermophilic enzymes has also been observed(Georlette et al., 2000). In this comparative thermostability study, putative ligase sequences retrieved from many hyperthermophilic and thermophilic environments such as HLH, TLA4, TLA5 and TLY have higher arginine residues in their primary structures compared to those from mesophilic and psychrophilic environments (table 5).

An exception was observed in *Candidatus Marinimicrobia bacterium* sp.1 group, where MLK (ligKDU4) has the highest Arg residues in its primary structure than its other pairs. Although, some ligase sequences from psychrophilic environments such as PLB show considerably high proline and arginine contents. Interestingly, some of these proline and arginine residues may not be at the strategic parts of the secondary structure (hot loops and exposed region) that aid the thermostability of the protein and this would be examined later.

## 5.2 Analysis of Ligase segments for disorder and globularity

Tight packing conformation and complex quaternary structure are characteristics of thermophiles (SUZUKI, 1989). With globularity, there is no general principle on globular protein thermostability. However, statistical analysis has shown that some aliphatic side chains (alanine,

valine, isoleucine and leucine) may increase the thermostability of globular proteins(Ikai, 1980; SUZUKI, 1989). An increase in globularity has also been linked to an increase in proline residues at the surface loops which protects the protein from thermal denaturation (Barzegar et al., 2009). All the ligase sequences examined in this comparative study have 75-100% of their amino acid residues in the globular regions with no general trend or rule in protein thermostability (table 6). Functional regions of protein lie within the globular regions and disordered regions(Linding, Russell, et al., 2003; Wright & Dyson, 1999).

The globular or ordered regions contain the regular secondary structures while intrinsically disordered regions (IDRs) do not have regular secondary structures, but show a higher degree of flexibility in the polypeptide chain (Linding, Jensen, et al., 2003). IDRs consist of a significant portion of both eukaryotic and prokaryotic proteomes (Burra et al., 2010). Many mammalian thermostable proteins have high intrinsic disordered regions (Galea et al., 2009). IDRs have shed more light on the adaptation of extremophiles to various extreme conditions; for instance, intrinsically disordered proteins withstand high heat treatment and often resistant to boiling temperatures (Burra et al., 2010; Kalthoff, 2003). Some intrinsically disordered proteins in plants are also cold-resistant, and that helps the plant to respond to water stress at the freezing temperatures (Kovacs et al., 2008; Tantos et al., 2009). These features support that IDRs promote adaptation to the extreme environment. However, there has been cross talk whether an increase in IDRs evolved as a strategy for thermal adaptation or they distinctively evolved since they are not part of globular or ordered regions (Burra et al., 2010; Thompson & Eisenberg, 1999). An opposing study has shown that psychrophilic and hyperthermophilic organisms may have average low protein disorder but a variable amount in mesophiles and thermophiles (Burra et al., 2010)

Although there is no general definition of protein disorder, according to the DisEMBL server, three different criteria were chosen for disordered regions, and that include loop/coils and hot loops. According to DisEMBL server, loops can be ordered or disordered and so did not meet all the requirement for being disorder, but the disordered regions in the loops are the non-helix and non  $\beta$ -strand states.(Linding, Jensen, et al., 2003) Hot loops are highly dynamic loops or loops with a high degree of mobility as defined by C-  $\alpha$  temperature factors (B-factors) and they have all the requirements of being disordered by the server(Kabsch & Sander, 1983; Linding, Jensen, et al., 2003) . In this comparative analysis, ligase sequences from the extreme environment (thermophilic and psychrophilic) have more amino residues in their disordered segments/hot loops than the mesophilic pairs under many groups, for instance, HLH (305), TLY (246) and PLB(182) under AB group have more residues in the disordered region(hot loops) than MLK2(table 6). Some ligase sequences from the psychrophilic environment, such as PLB2 and PLC2 have higher amino acid residues in the hot loop/disordered regions despite their shorter length residues (table 6).

**Table 6: Globularity and Disordered analysis of Ligase sequences**

Ligase.	Glo.Domains Res. (%)	Disordered Segments	
		Loops (Coils)Res.	Hot-loops Res
<b><i>Candidatus Marinimicrobia bacterium group Sp.1 (CBa group)</i></b>			
<b>HLG</b> (670)	668(98%)	431	274
<b>MLK</b> (630)	559(89%)	332	272
<b>PLC</b> (668)	596(89%)	354	276
<b><i>Acidimicrobiaceae bacterium group (AB group)</i></b>			
<b>HLH</b> (664)	497(75%)	407	305
<b>TLY</b> (681)	536(79%)	391	246
<b>MLK2</b> (567)	469(83%)	421	147
<i>E.coli</i> (671)	668(99%)	351	182
<b>PLB</b> (684)	642(93%)	414	182
<b>PLB2</b> (491)	471(96%)	313	195
<b><i>Moraxallaceae bacterium group (MB group)</i></b>			
<b>TLA4</b> (459)	457(99%)	304	81
<b>MLT2</b> (455)	455(100%)	168	71
<b>PLC2</b> (427)	425(99%)	230	131
<b><i>Phyllobacterium myrsinacearum group (PM group)</i></b>			
<b>TLA1</b> (714)	696(97%)	422	208
<b>TLA3</b> (541)	478(88%)	274	140
<b>MLT</b> (714)	698(98%)	418	216
<b><i>Rhizobiales bacterium group (RB group)</i></b>			
<b>TLA2</b> (637)	635(99%)	406	227
<i>E.coli</i> (671)	670(99%)	351	182
<b>PLM</b> (710)	707(99%)	383	244
<b><i>Candidatus marinimicrobia bacterium Sp.2(CBb group)</i></b>			
<b>HLG2</b> (398)	394(99%)	213	222
<b>TLA5</b> (444)	442(99%)	248	164
<b>MLK3</b> (397)	394(99%)	203	178
<b>MLK4</b> (338)	336(99%)	181	129
<b>PLS</b> (395)	392(99%)	154	106

\***Glo.Domains**(globular domains), **HLG(1&2)**: ligase sequences from the Hyperthermophilic hydrothermal Guayamas base sediment metagenomic assembly(Seitz et al., 2019), **HLH**: ligase sequence from the Hyperthermophilic Hot spring Iceland(Clum et al., 2009), **TLA**: Ligase sequences from Atlantis II, LCL metagenomic assembly **MLK**: ligase sequence from the Mesophilic Kebrit Deep upper surface metagenomic assembly, **MLT**: Ligase sequence from mesophilic terrestrial habitat(forest)(Tatusova et al., 2014), **PLC**: ligase sequence from the

Psychrophilic Cold, oxic seafloor metagenomic assembly(Tully, Wheat, et al., 2018), **TLY**: ligase sequence from moderately thermophilic mineral samples from the Yellowstone National Park United States(Tatusova et al., 2014), **PLB**: Ligase sequence from Psychrophilic Russia lake Baikal sub ice water metagenome(Cabello-Yeves et al., 2018), **PLB2**: Ligase sequence from Psychrophilic Japan lake Biwa freshwater metagenome(Hiraoka et al., 2019), **PLM**: ligase sequence from the Mediterranean sea, Marine Plankton Metagenomic assembly(Tully, Graham, et al., 2018). **PLS**: Ligase sequence from metagenome-assembled from Stordalen Mire thawing permafrost peatland(Martinez et al., 2019). Blue code for ligase sequences from hyperthermophilic environment, red code for thermophilic, orange for mesophilic and black for psychrophilic

### **5.3 Analysis of proline and arginine residues in the loops, hot loops, exposed and buried regions of the ligases**

Previous structural analyses of various alcohol dehydrogenases have revealed that prolines specifically in the hot loops or the constrained loops of thermophilic alcohol dehydrogenase are responsible for the increased thermostability over the mesophilic alcohol dehydrogenase (Barzegar et al., 2009; Bogin et al., 1998; Kelch & Agard, 2007; Sakaguchi et al., 2007). High frequency of proline within the loop region(hot loops) has been reported as a mechanism for thermostability adaptation (Barzegar et al., 2008; Georlette et al., 2000). Amino acid residues in the hot loops regions are rarely found in the buried/core regions but mostly in the exposed regions (Linding, Jensen, et al., 2003). Proline and arginine (PR) residues are either predicted to be in the loops and hot loops by DisEMBL server or exposed and buried regions by the ConSurf Server. We assume that PR residues would contribute to thermostability when proline or arginine residues are collaboratively predicted to be higher than other selected pairs at both the hot loops and exposed functional regions. In this comparative thermostability study, many ligase sequences coming from thermophilic or hyperthermophilic environments at different respective groups have higher number of proline and arginine residues both in the hot loops and exposed functional regions than their counterparts ligase sequences from mesophilic or psychrophilic environments(figure 16&17).

An exception was observed in the case of *MB* group, where the ligase sequence from the psychrophilic environment (PLC2) has the highest proline and arginine residues in the hot loops, but at the exposed or surface region it has the lowest proline residues (figure 16, Sup.fig 5). We do not have strong literature-based evidence for this, but we could suggest that some buried proline residues predicted by ConSuf server were predicted to be part of the hot loops residues by DisEMBL server, but the hot loops residues are mostly exposed(Linding, Jensen, et al., 2003). Two (TLA4&2) out of the five selected putative ligase sequences from the Atlantis II LCL layer have higher proline and arginine residues than their mesophilic and psychrophilic pairs in their respective groups(Figure 16 &17, Sup.fig 5, 6, 9 and 10). In the *MB* group, the ligase sequence from the mesophilic environment (MLT) has a similar number of proline and arginine residues both at the hot loops and exposed regions to one of the ligase sequence coming from the thermophilic Atlantis II LCL (TLA1)(figure 16&17, Sup.fig 7 and 8). In the *CBb* group, MLK 3 which was the only ligase sequence retrieved from the Kebrit deep lower surface layer has higher proline residue than MLK4 from Kebrit deep upper surfaces and other psychrophilic pair (figure 16, Sup.fig 11)

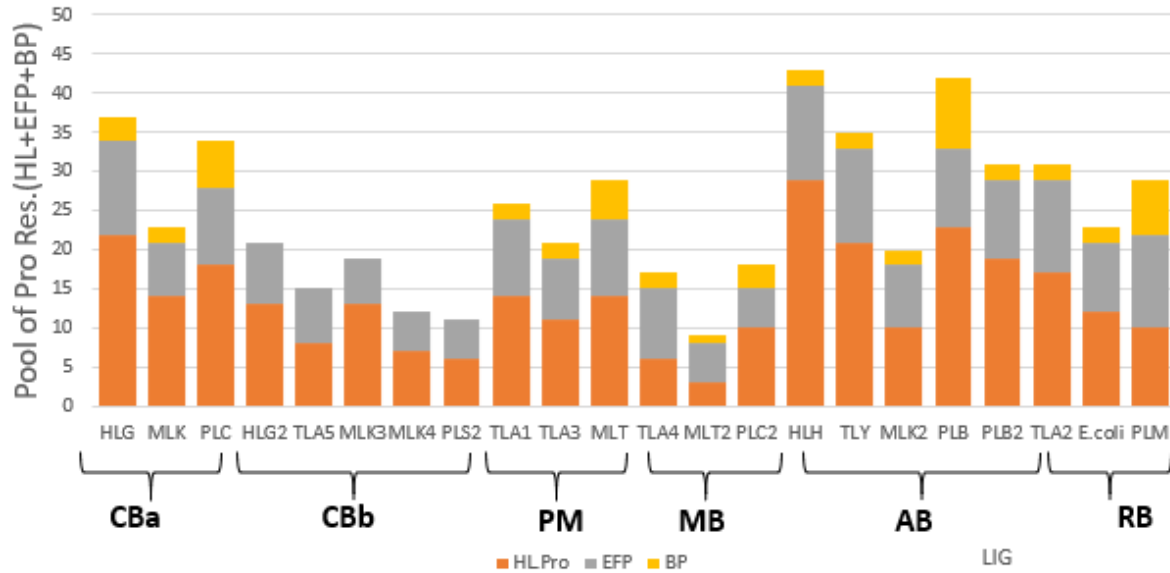


The position of the proline residues is also critical to enzyme thermostability in addition to its frequency. The insertion of proline residues at the external loop (exposed region) contributes more to alcohol dehydrogenase thermostability than any other region (Bogin et al., 1998). The loop proline residues reduce flexibility in the loop conformation, tighten the conformation packing, and maintain the enzyme's structure at higher temperatures (Barzegar et al., 2009). Proline is a unique amino acid because it has an imino acid group that is fixed at dihedral angle  $\phi$ , restricts the N-C rotation, and disrupts the tertiary protein conformation (Igarashi et al., 1999; Watanabe et al., 1991). Furthermore, proline also enhances the thermostability of protein by reducing the conformational entropy in addition to the restriction of the available conformational space caused by its pyrrolidine ring (Barzegar et al., 2009).

Arginine at the exposed region is frequent on the surfaces of many thermophilic proteins (Yokota et al., 2006). Increased thermostability has been achieved in many enzymes by mutating the lysine and hydrophobic amino acids at the exposed regions to arginine (Mortazavi & Hosseinkhani, 2011; Strub et al., 2004). In a previous study, the mutation of surface lysine to arginine improved protein stability in green fluorescent protein (Sokalingam et al., 2012). Also, an increased number of arginine residues in mesophilic over the thermophilic enzymes has also been reported (Georlette et al., 2000). In the case of this comparative thermostability study, arginine residues are more frequent in the hot loops and exposed regions of many ligase sequences from hyperthermophilic and thermophilic environments compared to other mesophilic and psychrophilic pairs across many groups. An exception was observed in the CBa group where MLK (ligKDU4) from the kebrit deep upper interface layer has the highest arginine residues in its group (Figure 17, Sup.fig 1). Although, this kind of exception has not been reported in literature but we could suggest that preferable amounts of Arg residues in the primary structure stay in the hot loop and exposed region.

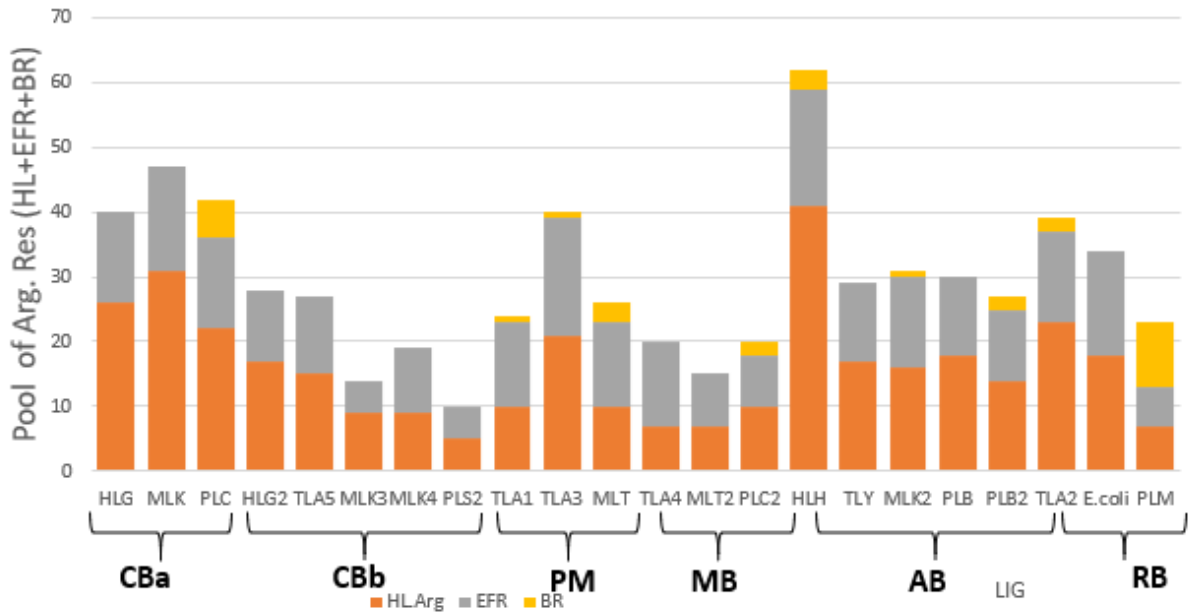
In addition, many ligase sequences retrieved from psychrophilic environments in this comparative thermostability study have the highest proline and arginine residues in the buried regions (figure 17&18). Amino acid substitutions and usage at the buried/core and exposed/surface regions are a common structural adaptation of protein to thermostability (Yokota et al., 2006). It has been hypothesized that psychrophilic enzymes have low stability and are able to maintain their activities at low temperatures due to an increase in global or local structural flexibility (Huston et al., 2008). Also, a decrease in proline residues and reduced accessibility of charge residues such as arginine may be responsible for high unfolding rate and low thermodynamic stability seen in cold-adapted enzymes (Huston et al., 2008). In the AB group, ligase sequences from psychrophilic environments (PLB) showed no lesser residues of proline and Arginine at the hot loop region, but it was observed that some of the proline and arginine hot loop residues were in the buried states thus, not contributing to thermostability (Figure 16& 17, Sup.fig 4) PLB has 23 pro residues in the hot loops' residues, 10 pro residues were predicted to be functionally exposed and 9 pro residues were predicted to be buried or even structurally buried. Also, in *Rhizobiales bacterium species group*, PLM has, in fact, more arginine residues as buried residues than functionally exposed residues (figure 17 and Sup.fig 9).

In summary, ligase sequences from hyperthermophilic and thermophilic environments have higher proline and arginine residues at strategic parts (hot loops and exposed regions) for structural protein thermostability. Ligase sequences from psychrophilic environments exhibit two properties for structural adaptation to non-thermostability. Firstly, a reduced number of proline and arginine residues predicted both in the hot loop and exposed functional regions, and secondly, reducing the accessibility of PR residues to the exposed or surface region (buried regions) (figure 16&17).



**Figure (16): Summary of Proline residue analysis at the hot loops, exposed functional and buried regions.** CBa is the *Candidatus marinimicrobia bacterium* species group, CBb is the *Candidatus marinimicrobia bacterium* (partial) species group, PM is the *Phyllobacterium myrsinacearum* species group, MB is the *Moraxallaceae bacterium* species group, AB is the *Acidimicrobiaceae bacterium* species group and RB is the *Rhizobiales bacterium* species group. HL.Pro Proline residues in the Hot-loops, EFP Exposed functional proline residues and BP is the Buried proline residues.



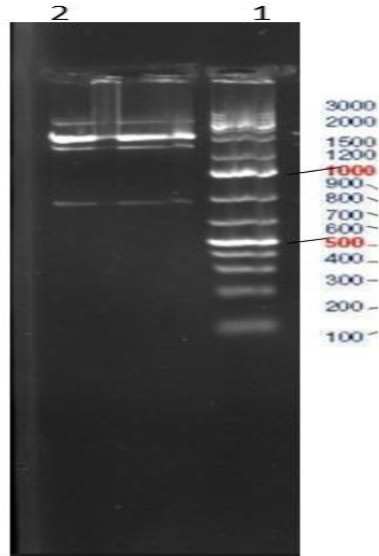


**Figure 17: Summary of Arginine residue analysis at the hot loops, Exposed functional and buried regions.** CBa is the *Candidatus marinimicrobia bacterium* species group, CBb is the *Candidatus marinimicrobia bacterium* (partial) species group, PM is the *Phyllobacterium myrsinacearum* species group, MB is the *Moraxallaceae bacterium* species group, AB is the *Acidimicrobiaceae bacterium* species group and RB is the *Rhizobiales bacterium* species group. HL.Arg arginine residues in the Hot-loops, EFR Exposed functional arginine residues and BP is the Buried arginine residues.

## 6. Gene Synthesis, Cloning and Transformation of LigATL1

### 6.1 Cloning and Transforming of LigATL1 to cloning host *E. coli* DH5 $\alpha$

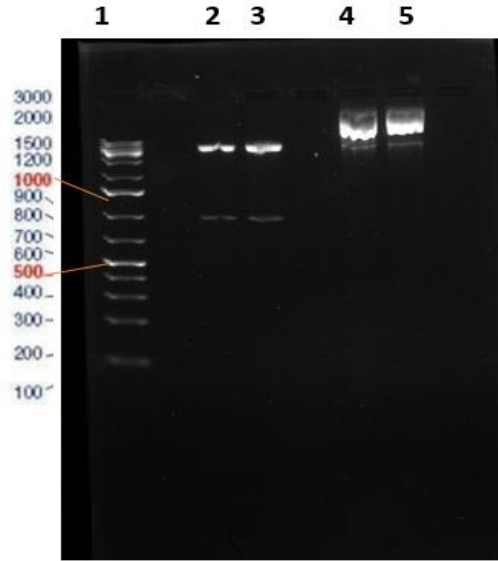
LigATL1 gene sequence from the Atlantis II LCL metagenomic datasets was first codon-optimized on the Gene Script Codon Optimization tool, modified for bi-directional cloning by BamHI and NdeI, and synthesized by Macrogen, Inc (South Korea). The ligase gene stored in PUC 19 was digested from the two restriction sites ends, NdeI and BamHI, and chemically transformed in competent *E. coli* DH5 $\alpha$ . The Plasmid DNA extracted from the prepared overnight culture of a positively transformed clone of *E. coli* DH5  $\alpha$  was digested by the same restriction enzymes using the NEB standard protocol (figure 18). The ligase gene was gel purified by Zymoclean<sup>TM</sup> Gel DNA Recovery Kit (Zymo Research, Irvine, CA, USA) and the DNA concentration estimated by NanoDrop<sup>TM</sup> 2000/2000c Spectrophotometer.



**Figure (18) : Restriction digestion of recombinant pUC19 containing ligase gene insert at NdeI and Bam HI sites.** Lane 1; ; GeneRuler<sup>TM</sup> 100bp plus DNA ladder (Thermoscientific), lane 2; pUC 19 digested with NdeI and Bam HI.

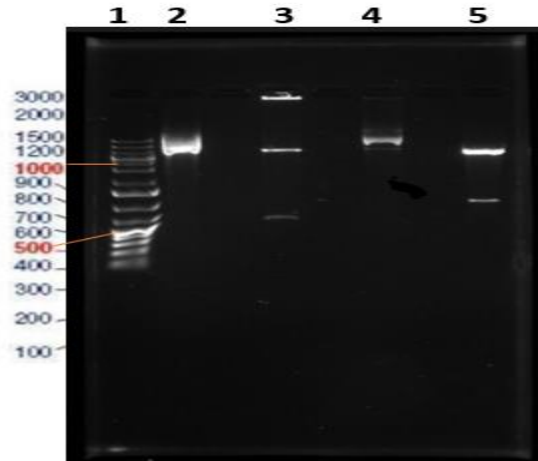
## 6.2 Cloning and Transforming of LigATLI in Expression hosts

The ligase gene insert was digested from pUC 19 and ligated with an expression vector, pET-16b (Novagen), following the NEB standard protocol. The ligation was confirmed on 1% agarose gel, and after which the recombinant pET 16b was chemically transformed into *E coli* BL21(DE3) and *E coli* PlyS (Chang et al., 2017).



**Figure (19): Restriction digestion of recombinant pET 16B containing ligase gene insert.** The lane 1; GeneRuler™ 100bp plus DNA ladder (Thermoscientific), lane 2&3; recombinant pET 16B digested with NdeI and Bam HI, lane 5 & 6; undigested pET16B.

The size of the correct frame of the gene was confirmed by digesting with the restriction enzyme, KpnI which has a site within only the insert, but not in the holding vector (pET-16 b) (figure 20).



**Figure (20): Restriction digestion of recombinant pET 16b containing ligase gene insert to confirm orientation of the insert.** The lane 1; GeneRuler™ 100bp plus DNA ladder (Thermoscientific), lane 2; Recombinant pET 16B without digestion, lane 3; pET 16B digested with KpnI + NdeI, lane 4; pET 16B digested with KpnI only lane5 ; pE 16B digested with BamHI and NdeI

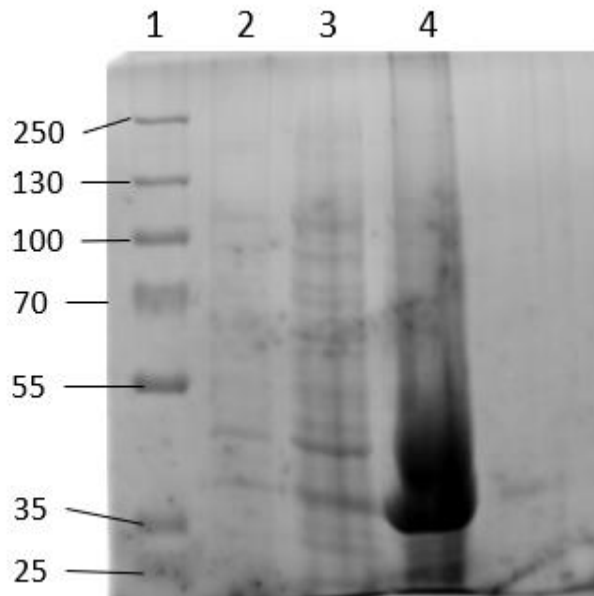
### 6.3 Expression of Lig-ATL1

pET 16B was used as the expression vector. It carries 10X N terminal His-Tag, three cloning sites and T7 expression region that is transcribed by T7 RNA polymerase. The protein product of LigATL1 as N-terminal-10- histidine-tagged fusion recombinant protein was predicted to be 48kDa due to the extra 11kDa and 3kDa of pET-fused protein and 10-Histidine tag, respectively. This expression vector enhanced solubility as well as increase the solubility of recombinant ligase protein. Shuman et al., 2005 have previously used pET 16B to express ATP-dependent DNA ligase from euryarchaeon *Pyrococcus horikoshii* (Keppetipola & Shuman, 2005) under an induction time of 6hrs and IPTG concentration of 0.4mM. LigATL1-pET 16B expression was initially induced following the same condition in *E coli* BL21(DE3), but a large proportion of the protein went into inclusion bodies (figure 21). The induction time was reduced to 3hrs due to aggressive smear noticed on the gel, but no protein was induced both in *E coli* BL21 DE3 and *E. coli* BL21 PlyS (figure 22). We concluded on optimizing with the previous condition of Shuman et al., 2005.

Various optimizations conditions (change of lysis buffer pH, NaCl concentration, IPTG concentration and expression time) were tried before arriving at a good expression and solubility. We decided to first optimize the solubility of the protein expression by adjusting the lysis buffer (Sodium phosphate buffer, pH7.4, 0.5M NaCl) pH to be at least 2 units away from the Expasy calculated theoretical pI of the LigATL1 which is 7.59 because the previous experiments were conducted using the same lysis buffer. This approach of pI difference to solve insolubility in protein has been used by a member of our research team to successfully express transglycosylase

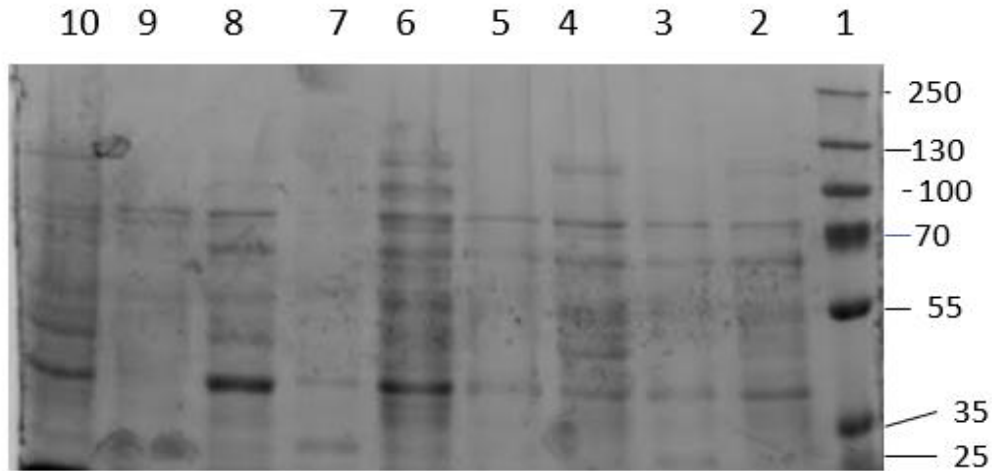
from an extremophilic environment (Kirkwood et al., 2015; Malash et al., 2020; Zhang et al., 2013).

Two lysis buffer of acidic pH 6.4 and basic pH 8.5 which are within the buffering change of Sodium phosphate buffer were chosen. pH 6.4 lysis buffer did not put the recombinant protein in the soluble fraction, but some little portion of the recombinant protein was present in the soluble fraction with the sodium-phosphate lysis buffer pH 8.5 (figure 23). Unfortunately, under the large-scale expression, all the protein fractions went into insolubility with the same condition because all the previous conditions were tried on the small-scale level. We decided to try two things, to recalculate the pI of the protein using a different tool and to use different lysis buffer (Sodium glycine buffer). Another pI of 6.98 which was calculated on isoelectric.org, was obtained for LigATL1, and this prompted the adjustment of the Sodium lysis buffer to pH 8.0, following the manufacturer's instruction(GE Healthcare) suggestion if the protein was insoluble at pH 7.4. We also decided to reduce the induction time to 5hrs and salt concentration to 0.3M NaCl due to the aggressive smear noticed on the gel, and the fact that some proteins are always soluble at very low salt concentration(Inouye et al., 1998). The protein was found to be sufficiently soluble at Sodium-phosphate buffer pH 8.0, 0.3M NaCl, IPTG concentration of both 0.1 and 0.4mM with the induction time of 5hrs (figure 24). The same previous condition was also successful at the large-scale expression (figure 25).



**Figure (21) : SDS-PAGE Analysis of LigATL1 following the expression in E.coli BL21 DE3 induction condition of 6hrs at 37 °C, lysis buffer pH 7.4, and IPTG conc. 0.4mM and 0.5mM NaCl**

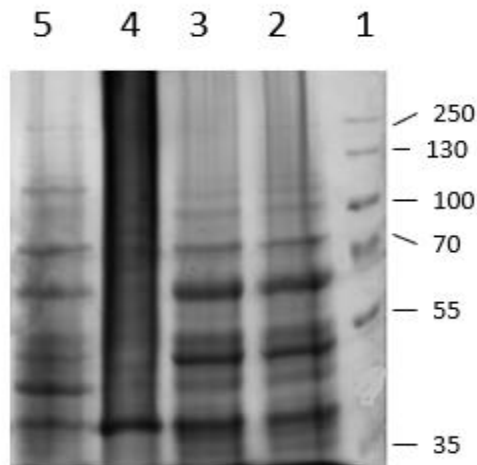
Lane 1; ThermoScientific™ Colour Protein Ladder( 10-250kDa), lane 2; Uninduced sample (cell lysate before inducing with IPTG), lane 3; Soluble fraction (supernatant), lane 4; Pellet (Cell debris).



**Figure (22): SDS-PAGE Analysis of LigATL1 following the expression in E.coli BL21 DE3 and E.coli PLYS induction condition of 3hrs at 37 °C and overnight (16hrs at 16°C), lysis buffer pH 7.4, IPTG conc. 0.4mM and 0.5mM NaCl**

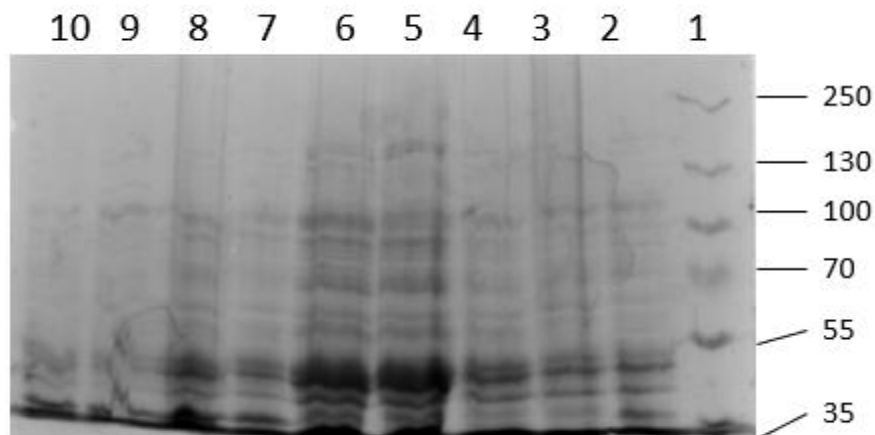
Lane 1; ThermoScientific™ Colour Protein Ladder( 10-250kDa), lane 2; Un-induced sample( cell lysate before IPTG induction), lane 3; Soluble fraction at 3hrs induction for E. coli BL21 DE3, lane 4; Insoluble fraction at 3hrs induction, E.coli BL21 DE3, lane 5; Soluble fraction at overnight induction for E coli BL21 DE3, lane 6; Insoluble fraction at overnight induction for E coli BL21 DE3, lane 7; Soluble fraction at overnight induction for E coli BL21 PLYS, lane 8; Insoluble fraction at overnight induction for E coli BL21 PLYS, lane 9; Soluble fraction at 3hrs induction for E. coli BL21 PLYS, lane 10; Insoluble fraction at 3hrs induction, E.coli BL21 PLYS.





**Figure(23): SDS-PAGE Analysis of LigATL1 following the expression in *E.coli* BL21 DE3, induction condition of 6hrs, lysis buffer pH 8.5 &6.4, IPTG conc. 0.4mM and 0.5mM NaCl**

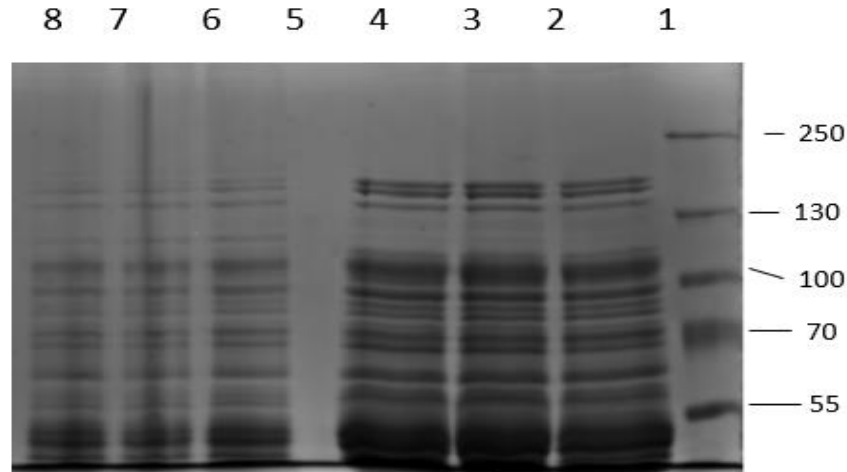
Lane 1;Thermoscientific™ Colour Protein Ladder( 10-250kDa), lane 2; Insoluble fraction (cell debris) at lysis buffer pH 8.5, lane 3; Soluble fractions(supernatant) at lysis buffer pH 8.5, lane 4; ; Insoluble fraction (cell debris) at lysis buffer pH 6.4, lane 5; Soluble fractions(supernatant) at lysis buffer pH 6.4.



**Figure(24) : SDS-PAGE Analysis of LigATL1 following the expression in *E.coli* BL21 DE3, induction condition of 5hrs using different lysis buffers( Na-Glycine buffer pH 10.0, Na-Phosphate buffer pH 8.0), different IPTG conc( 0.1& 0.4mM) and 0.3mM NaCl**

Lane1; Thermoscientific™ Colour Protein Ladder( 10-250kDa), lane 2; Un-induced sample( cell lysate before adding IPTG), lane3; Soluble fraction at Na-Gly Buffer, pH 10.0, IPTG 0.4mM, lane 4;Soluble fraction at Na-Gly Buffer, pH 10.0, IPTG 0.1mM, lane 5; Soluble fraction at Na-Phosphate Buffer, pH 8.0, IPTG 0.1mM, lane 6; Soluble fraction at Na-Phosphate Buffer, pH 8.0,

IPTG 0.4mM, lane 7; Insoluble fraction at Na-Gly Buffer, pH 10.0, IPTG 0.4mM, lane 8; Insoluble fraction at Na-Gly Buffer, pH 10.0, IPTG 0.1mM, lane 9; Insoluble fraction at Na-Phosphate Buffer, pH 8.0, IPTG 0.4mM, lane 10; Insoluble fraction at Na-Phosphate Buffer, pH 8.0, IPTG 0.1mM



**Figure(25): SDS-PAGE Analysis of LigATL1 following the expression in *E.coli BL21* DE3, induction condition of 5hrs using Na-Phosphate lysis buffer pH 8.0, IPTG 0.1mM and 0.3mM NaCl at large scale expression.**

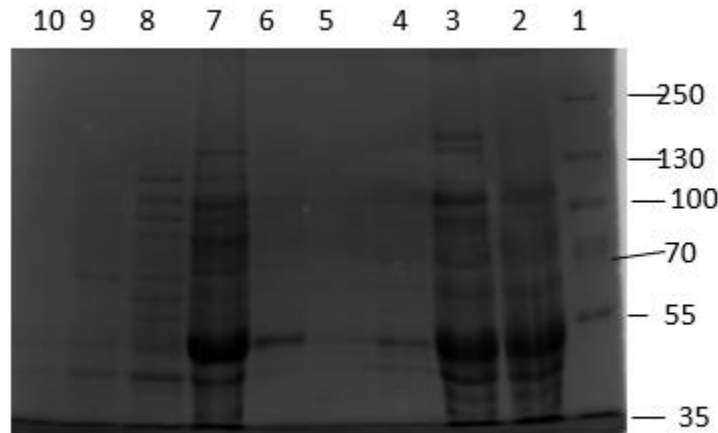
Lane 1; ThermoScientific™ Colour Protein Ladder( 10-250kDa), lane 2-4 ; soluble fraction ( supernatant), lane 6-8; insoluble fraction ( cell debris).

#### 6.4 Partial Purification of His-tagged Lig-ATL1 on Ni<sup>2+</sup> column

LigATL1 which is N-terminal His-tagged protein was initially tried for purification with Ni<sup>2+</sup> ions immobilized on NTA column matrix. The binding buffer contained 20mM of NaH<sub>2</sub> PO<sub>4</sub>, pH 8.0 0.3M NaCl, 10% glycerol, 0.2% Triton X and 20mM imidazole, and the Elution buffer has the same components except that the imidazole concentration was 500mM which is according to the manufacturer's instruction. Unfortunately, no purified protein was observed on the SDS gel, and we instantly suspected that the imidazole concentration was high and that it completely inhibited the protein from binding to the Ni<sup>2+</sup> column (data not shown). We reduced the imidazole concentration in the binding buffer in a stepwise manner (15-0mM) and only in 0mM imidazole that the protein-bound effectively (data not shown). We assumed that the protein could only bind to the metal column without imidazole. Another purification trial was done where the His-tagged protein was stepped wisely eluted with Elution buffers having imidazole concentration ranging from 10-500mM in order to optimize the imidazole in the washing buffer (figure 26). The His-tagged protein eluted with just 30mM imidazole before many contaminants were eluted. We decided to elute with the only 30mM of imidazole in a stepwise manner, and only then, we were

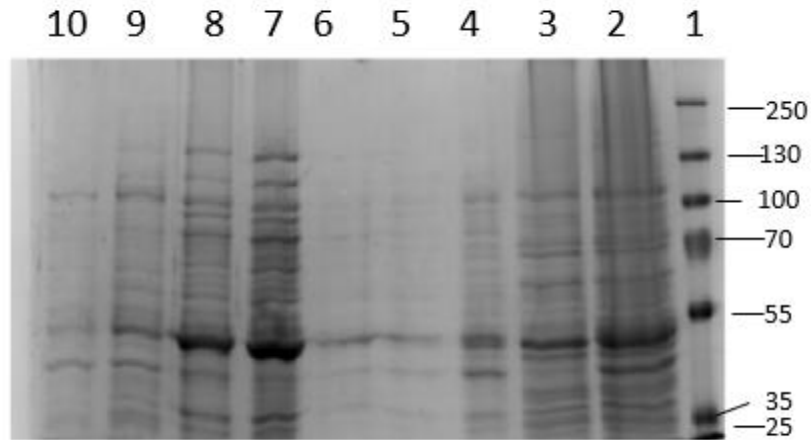
able to purify the protein partially (figure 27). All the purification trials were done under the native conditions because the refolding of many extremophilic proteins have failed to produce soluble and active enzymes for preliminary functional tests(Malash et al., 2020).

One possible explanation for the reduced binding capacity of His-tagged protein in the presence of imidazole is that the N-terminal His-tag is not totally accessible for binding on the column probably the tag might have been covered slightly by the protein folding. In this case, the His-tag could be moved to the C-terminus of the vector to improve the binding capacity of the His-tagged protein. Some other suggestions include re-cloning the LigATL1 into another vector that uses both the GST(Glutathione S-transferase) tag and His tag such as pETM30 vector (EMBL) or by adding a flexible linker between His-tag and the protein(Z. Li et al., 2011). A two-step purification procedure (affinity and ion-exchange chromatography) may also help achieve the full purification of the protein.



**Figure (26): SDS-PAGE analysis of partial purification of LigATL1 on Ni-NTA column under the native condition with binding buffer (20mM of  $\text{NaH}_2\text{PO}_4$ , pH 8.0, 0.3M NaCl, 10% glycerol, 0.2% Triton X and 0mM imidazole) and imidazole stepwise elution (10-500mM)**

Lane1;Thermoscientific <sup>TM</sup> Colour Protein Ladder( 10-250kDa), lane 2; Soluble extract(supernatant) lane 3; flowthrough lane 4; Wash, lane 5; Elute with 10mM imidazole, lane 6; Elute with 30mM imidazole, lane 7; Elute with 50mM imidazole, lane 8; Elute with 70mM imidazole, lane 9; Elute with 100mM imidazole, lane 10; Elute with 500mM imidazole.



**Figure(27): SDS-PAGE analysis of partial purification of LigATL on Ni-NTA column under native condition with binding buffer (20mM of  $\text{NaH}_2\text{PO}_4$ , pH 8.0, 0.3M NaCl, 10% glycerol, 0.2% Triton X and 0mM imidazole) and stepwise elution from 30-40mM.**

Lane1; ThermoScientific™ Colour Protein Ladder( 10-250kDa), lane2; flowthrough, lane3; wash, lane4; Elute with 20mM imidazole, lane 5-8; Elute with 30mM imidazole, lane 9; Elute with 35mM imidazole, lane 10; Elute with 40mM imidazole.

## Chapter 4: Conclusion and Recommendation

DNA ligase has been an invaluable enzyme to biotechnology, molecular biology and biomedicine ever since its discovery in 1967. Thermostable ATP-dependent DNA ligase forms the bedrock of Ligase Chain reaction (LCR) which is used in the molecular detection of various pathogens and diseases while NAD<sup>+</sup>-dependent DNA ligase has been incorporated in chemotherapy for drug designs. The available metagenomic database has made the mining and exploration of extreme novel biocatalysts from extremophilic environments such as the Red Sea environment possible.

In this study, 18 ORFs for putative DNA ligases were identified, and two of the ORFs LigATL1 (ATP type) and LigKDU4 (NAD<sup>+</sup> type) that have low identities to the ligases in the NCBI database were synthesized. The phylogenetic analysis shows that LigATL1 belongs to LigD family while LigKDU4 is present in the LigA family. The Preliminary characterizations (3D modeling and superimposition predictions) revealed that LigATL1 and LigKDU4 sequences modeled with 100% confidence and superimposed accurately with thermostable DNA ligases.

LigATL1 was further cloned, expressed and partially purified. However, to achieve better purity with the protein, two steps purification may be done. The binding capacity of the protein may be improved by recloning LigATL1 in another vector such as pETM30 vector(EMBL), or reconstructing the previous vector to include a flexible linker between the tag and protein(Z. Li et al., 2011). Also, *In vitro* functional assay and biochemical characterizations such as examining the effect of pH, temperature and salts are required to evaluate the activity and stability of the enzyme to better exploit its potentials in the biotechnological application.

In the comparative thermostability study, some ligase sequences from thermophilic environments such as the Atlantis II LCL layer has higher exposed proline and arginine residues at the strategic loop area(hot loops) as a means of thermal adaptation while those from psychrophilic environments buried some of their proline and arginine residues to exhibit a non-thermostability phenomenon. Higher proline residues at the hot loops of thermophilic alcohol dehydrogenase have also been previously reported(Barzegar et al., 2009). Thermostable Luciferase has been previously designed by substituting hydrophobic amino acids to arginine at the exposed regions(Mortazavi & Hosseinkhani, 2011). The hot loop and exposed region analyses can be exploited to enhance the thermal stability of mesophilic enzymes by specifically mutating other amino residues such as lysine and hydrophobic amino acids to proline and arginine residues at these loop areas.

## References

- Abdallah, R. Z., Adel, M., Ouf, A., Sayed, A., Ghazy, M. A., Alam, I., Essack, M., Lafi, F. F., Bajic, V. B., & El-Dorry, H. (2014). Aerobic methanotrophic communities at the Red Sea brine-seawater interface. *Frontiers in Microbiology*, *5*, 487.
- Abida, H., Ruchaud, S., Rios, L., Humeau, A., Probert, I., De Vargas, C., Bach, S., & Bowler, C. (2013). Bioprospecting marine plankton. *Marine Drugs*, *11*(11), 4594–4611.
- Anschutz, P., Blanc, G., Chatin, F., Geiller, M., & Pierret, M.-C. (1999). Hydrographic changes during 20 years in the brine-filled basins of the Red Sea. *Deep Sea Research Part I: Oceanographic Research Papers*, *46*(10), 1779–1792.
- Antunes, A., Ngugi, D. K., & Stingl, U. (2011). Microbiology of the Red Sea (and other) deep-sea anoxic brine lakes. *Environmental Microbiology Reports*, *3*(4), 416–433.
- Arrieta, J. M., Arnaud-Haond, S., & Duarte, C. M. (2010). What lies underneath: Conserving the oceans' genetic resources. *Proceedings of the National Academy of Sciences*, *107*(43), 18318–18324.
- Ashkenazy, H., Erez, E., Martz, E., Pupko, T., & Ben-Tal, N. (2010). ConSurf 2010: Calculating evolutionary conservation in sequence and structure of proteins and nucleic acids. *Nucleic Acids Research*, *38*(suppl\_2), W529–W533.
- Backer, H., & Schoell, M. (1972). New deeps with brines and metalliferous sediments in the Red Sea. *Nature Physical Science*, *240*(103), 153–158.
- Bailey, J. M., Lin, L.-N., Brandts, J. F., & Mas, M. T. (1990). Substitution of a proline for alanine 183 in the hinge region of phosphoglycerate kinase: Effects on catalysis, activation by sulfate, and thermal stability. *Journal of Protein Chemistry*, *9*(1), 59–67.
- Bang, D., & Church, G. M. (2008). Gene synthesis by circular assembly amplification. *Nature Methods*, *5*(1), 37–39.



- Barany, F. (1991). Genetic disease detection and DNA amplification using cloned thermostable ligase. *Proceedings of the National Academy of Sciences*, 88(1), 189–193.
- Barone, R., De Santi, C., Palma Esposito, F., Tedesco, P., Galati, F., Visone, M., Di Scala, A., & De Pascale, D. (2014). Marine metagenomics, a valuable tool for enzymes and bioactive compounds discovery. *Frontiers in Marine Science*, 1, 38.
- Barzegar, A., Moosavi-Movahedi, A. A., Mahnam, K., Bahrami, H., & Sheibani, N. (2008). Molecular dynamic simulations of nanomechanic chaperone peptide and effects of in silico His mutations on nanostructured function. *Journal of Peptide Science: An Official Publication of the European Peptide Society*, 14(11), 1173–1182.
- Barzegar, A., Moosavi-Movahedi, A. A., Pedersen, J. Z., & Miroliaei, M. (2009). Comparative thermostability of mesophilic and thermophilic alcohol dehydrogenases: Stability-determining roles of proline residues and loop conformations. *Enzyme and Microbial Technology*, 45(2), 73–79.
- Bates, G. (2003). Huntingtin aggregation and toxicity in Huntington's disease. *The Lancet*, 361(9369), 1642–1644.
- Behzad, H., Ibarra, M. A., Mineta, K., & Gojobori, T. (2016). Metagenomic studies of the Red Sea. *Gene*, 576(2), 717–723.
- Bhanjdeo, M. M., Nayak, A. K., & Subudhi, U. (2017). Surface-assisted DNA self-assembly: An enzyme-free strategy towards formation of branched DNA lattice. *Biochemical and Biophysical Research Communications*, 485(2), 492–498.
- Blanc, G., & Anschutz, P. (1995). New hydrographic situation in the Atlantis II Deep hydrothermal brine system. *Geology*, 23, 543–546.

- Bogin, O., Peretz, M., Hacham, Y., Burstein, Y., Korkhin, Y., Kalb, A. J., & Frolow, F. (1998). Enhanced thermal stability of *Clostridium beijerinckii* alcohol dehydrogenase after strategic substitution of amino acid residues with prolines from the homologous thermophilic *Thermoanaerobacter brockii* alcohol dehydrogenase. *Protein Science*, 7(5), 1156–1163.
- Bowater, R. P., Cobb, A. M., Pivonkova, H., Havran, L., & Fojta, M. (2015). Biophysical and electrochemical studies of protein–nucleic acid interactions. *Monatshefte Für Chemie-Chemical Monthly*, 146(5), 723–739.
- Burra, P. V., Kalmar, L., & Tompa, P. (2010). Reduction in structural disorder and functional complexity in the thermal adaptation of prokaryotes. *PloS One*, 5(8).
- Cabello-Yeves, P. J., Zenskaya, T. I., Rosselli, R., Coutinho, F. H., Zakharenko, A. S., Blinov, V. V., & Rodriguez-Valera, F. (2018). Genomes of novel microbial lineages assembled from the sub-ice waters of Lake Baikal. *Appl. Environ. Microbiol.*, 84(1), e02132–17.
- Chang, A. Y., Chau, V. W., Landas, J. A., & Pang, Y. (2017). Preparation of calcium competent *Escherichia coli* and heat-shock transformation. *JEMI Methods*, 1, 22–25.
- Chao, Y.-C., Merritt, M., Schaefferkoetter, D., & Evans, T. G. (2020). High-throughput quantification of protein structural change reveals potential mechanisms of temperature adaptation in *Mytilus* mussels. *BMC Evolutionary Biology*, 20(1), 1–18.
- Clum, A., Nolan, M., Lang, E., Del Rio, T. G., Tice, H., Copeland, A., Cheng, J.-F., Lucas, S., Chen, F., & Bruce, D. (2009). Complete genome sequence of *Acidimicrobium ferrooxidans* type strain (ICP T). *Standards in Genomic Sciences*, 1(1), 38.
- Cochran, J. R. (1983). A model for development of Red Sea. *Aapg Bulletin*, 67(1), 41–69.

- Cole, S. T., Eiglmeier, K., Parkhill, J., James, K. D., Thomson, N. R., Wheeler, P. R., Honore, N., Garnier, T., Churcher, C., & Harris, D. (2001). Massive gene decay in the leprosy bacillus. *Nature*, *409*(6823), 1007–1011.
- Cronan, D. S. (1999). *Handbook of marine mineral deposits* (Vol. 18). CRC press.
- D'Argenio, V., & Salvatore, F. (2015). The role of the gut microbiome in the healthy adult status. *Clinica Chimica Acta*, *451*, 97–102.
- Degens, E. T., & Ross, D. A. (2013). *Hot brines and recent heavy metal deposits in the Red Sea: A geochemical and geophysical account*. Springer-Verlag.
- Doherty, A. J., & Suh, S. W. (2000). Structural and mechanistic conservation in DNA ligases. *Nucleic Acids Research*, *28*(21), 4051–4058.
- Dunker, A. K., Lawson, J. D., Brown, C. J., Williams, R. M., Romero, P., Oh, J. S., Oldfield, C. J., Campen, A. M., Ratliff, C. M., & Hipps, K. W. (2001). Intrinsically disordered protein. *Journal of Molecular Graphics and Modelling*, *19*(1), 26–59.
- Dwivedi, N., Dube, D., Pandey, J., Singh, B., Kukshal, V., Ramachandran, R., & Tripathi, R. P. (2008). NAD<sup>+</sup>-Dependent DNA Ligase: A novel target waiting for the right inhibitor. *Medicinal Research Reviews*, *28*(4), 545–568.
- Eder, W., Jahnke, L. L., Schmidt, M., & Huber, R. (2001). Microbial diversity of the brine-seawater interface of the Kebrit Deep, Red Sea, studied via 16S rRNA gene sequences and cultivation methods. *Applied and Environmental Microbiology*, *67*(7), 3077–3085.
- Eder, W., Schmidt, M., Koch, M., Garbe-Schönberg, D., & Huber, R. (2002). Prokaryotic phylogenetic diversity and corresponding geochemical data of the brine-seawater interface of the Shaban Deep, Red Sea. *Environmental Microbiology*, *4*(11), 758–763.

- Eggerding, F., winndeen, e., giusti, w., adriano, t., iovannisci, d., & brinson, e. (1993). detection of mutations in the cystic-fibrosis gene by multiplex amplification and oligonucleotide ligation. *american journal of human genetics*, 53, 1485–1485.
- Eijsink, V. G., Bjørk, A., G\aaaseidnes, S., Sirev\aaag, R., Synstad, B., van den Burg, B., & Vriend, G. (2004). Rational engineering of enzyme stability. *Journal of Biotechnology*, 113(1–3), 105–120.
- Elbehery, A. H., Leak, D. J., & Siam, R. (2017). Novel thermostable antibiotic resistance enzymes from the Atlantis II Deep Red Sea brine pool. *Microbial Biotechnology*, 10(1), 189–202.
- Emerson, D., Rentz, J. A., Lilburn, T. G., Davis, R. E., Aldrich, H., Chan, C., & Moyer, C. L. (2007). A novel lineage of proteobacteria involved in formation of marine Fe-oxidizing microbial mat communities. *PloS One*, 2(8), e667.
- Erlich, H. A., Gelfand, D., & Sninsky, J. J. (1991). Recent advances in the polymerase chain reaction. *Science*, 252(5013), 1643–1651.
- Faber, E., Botz, R., Poggenburg, J., Schmidt, M., Stoffers, P., & Hartmann, M. (1998). Methane in Red Sea brines. *Organic Geochemistry*, 29(1–3), 363–379.
- Fields, P. A. (2001). Protein function at thermal extremes: Balancing stability and flexibility. *Comparative Biochemistry and Physiology Part A: Molecular & Integrative Physiology*, 129(2–3), 417–431.
- Gajiwala, K. S., & Pinko, C. (2004). Structural rearrangement accompanying NAD<sup>+</sup> synthesis within a bacterial DNA ligase crystal. *Structure*, 12(8), 1449–1459.
- Galea, C. A., High, A. A., Obenauer, J. C., Mishra, A., Park, C.-G., Punta, M., Schlessinger, A., Ma, J., Rost, B., & Slaughter, C. A. (2009). Large-scale analysis of thermostable,

- mammalian proteins provides insights into the intrinsically disordered proteome. *Journal of Proteome Research*, 8(1), 211–226.
- Gallagher, S. R. (2006). One-dimensional SDS gel electrophoresis of proteins. *Current Protocols in Molecular Biology*, 75(1), 10–2.
- Georlette, D., Jonsson, Z. O., Van Petegem, F., Chessa, J.-P., Van Beeumen, J., Hübscher, U., & Gerday, C. (2000). A DNA ligase from the psychrophile *Pseudoalteromonas haloplanktis* gives insights into the adaptation of proteins to low temperatures. *European Journal of Biochemistry*, 267(12), 3502–3512.
- Gibriel, A. A., & Adel, O. (2017). Advances in ligase chain reaction and ligation-based amplifications for genotyping assays: Detection and applications. *Mutation Research/Reviews in Mutation Research*, 773, 66–90.
- Gibson, D. G. (2011). Enzymatic assembly of overlapping DNA fragments. In *Methods in enzymology* (Vol. 498, pp. 349–361). Elsevier.
- Goihberg, E., Dym, O., Tel-Or, S., Levin, I., Peretz, M., & Burstein, Y. (2007). A single proline substitution is critical for the thermostabilization of *Clostridium beijerinckii* alcohol dehydrogenase. *Proteins: Structure, Function, and Bioinformatics*, 66(1), 196–204.
- Goldenzweig, A., & Fleishman, S. J. (2018). Principles of protein stability and their application in computational design. *Annual Review of Biochemistry*, 87, 105–129.
- Gu, W., Wang, T., Maltais, F., Ledford, B., Kennedy, J., Wei, Y., Gross, C. H., Parsons, J., Duncan, L., & Arends, S. R. (2012). Design, synthesis and biological evaluation of potent NAD<sup>+</sup>-dependent DNA ligase inhibitors as potential antibacterial agents. Part I: Aminoalkoxypyrimidine carboxamides. *Bioorganic & Medicinal Chemistry Letters*, 22(11), 3693–3698.

- Handelsman, J. (2004). Metagenomics: Application of genomics to uncultured microorganisms. *Microbiol. Mol. Biol. Rev.*, 68(4), 669–685.
- Hartmann, M., Scholten, J. C., Stoffers, P., & Wehner, F. (1998). Hydrographic structure of brine-filled deeps in the Red Sea—New results from the Shaban, Kebrit, Atlantis II, and Discovery Deep. *Marine Geology*, 144(4), 311–330.
- Hawkins, T. L., O'Connor-Morin, T., Roy, A., & Santillan, C. (1994). DNA purification and isolation using a solid-phase. *Nucleic Acids Research*, 22(21), 4543.
- He, X., Ni, X., Wang, Y., Wang, K., & Jian, L. (2011). Electrochemical detection of nicotinamide adenine dinucleotide based on molecular beacon-like DNA and E. coli DNA ligase. *Talanta*, 83(3), 937–942.
- Hiraoka, S., Okazaki, Y., Anda, M., Toyoda, A., Nakano, S., & Iwasaki, W. (2019). Metaepigenomic analysis reveals the unexplored diversity of DNA methylation in an environmental prokaryotic community. *Nature Communications*, 10(1), 1–10.
- Huang, Y.-F., Chen, S.-C., Chiang, Y.-S., Chen, T.-H., & Chiu, K.-P. (2012). Palindromic sequence impedes sequencing-by-ligation mechanism. *BMC Systems Biology*, 6, S10.
- Huber, H., & Stetter, K. O. (2015). Deferribacteres class. Nov. *Bergey's Manual of Systematics of Archaea and Bacteria*, 1–1.
- Huson, D. H., Auch, A. F., Qi, J., & Schuster, S. C. (2007). MEGAN analysis of metagenomic data. *Genome Research*, 17(3), 377–386.
- Huston, A. L., Haeggström, J. Z., & Feller, G. (2008). Cold adaptation of enzymes: Structural, kinetic and microcalorimetric characterizations of an aminopeptidase from the Arctic psychrophile *Colwellia psychrerythraea* and of human leukotriene A4 hydrolase. *Biochimica et Biophysica Acta (BBA)-Proteins and Proteomics*, 1784(11), 1865–1872.



- Iakoucheva, L. M., Brown, C. J., Lawson, J. D., Obradović, Z., & Dunker, A. K. (2002). Intrinsic disorder in cell-signaling and cancer-associated proteins. *Journal of Molecular Biology*, 323(3), 573–584.
- Igarashi, K., Ozawa, T., Ikawa-Kitayama, K., Hayashi, Y., Araki, H., ENDO, K., Hagihara, H., Ozaki, K., Kawai, S., & Ito, S. (1999). Thermostabilization by proline substitution in an alkaline, liquefying  $\alpha$ -amylase from *Bacillus* sp. Strain KSM-1378. *Bioscience, Biotechnology, and Biochemistry*, 63(9), 1535–1540.
- Ikai, A. (1980). Thermostability and aliphatic index of globular proteins. *The Journal of Biochemistry*, 88(6), 1895–1898.
- Inouye, K., Kuzuya, K., & Tonomura, B. (1998). Effect of salts on the solubility of thermolysin: A remarkable increase in the solubility as well as the activity by the addition of salts without aggregation or dispersion of thermolysin. *The Journal of Biochemistry*, 123(5), 847–852.
- Jeon, H. J., Shin, H.-J., Choi, J. J., Hoe, H.-S., Kim, H.-K., Suh, S. W., & Kwon, S.-T. (2004). Mutational analyses of the thermostable NAD<sup>+</sup>-dependent DNA ligase from *Thermus filiformis*. *FEMS Microbiology Letters*, 237(1), 111–118.
- Jia, B., Xuan, L., Cai, K., Hu, Z., Ma, L., & Wei, C. (2013). NeSSM: A next-generation sequencing simulator for metagenomics. *PloS One*, 8(10), e75448.
- Kabsch, W., & Sander, C. (1983). Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers: Original Research on Biomolecules*, 22(12), 2577–2637.
- Kalthoff, C. (2003). A novel strategy for the purification of recombinantly expressed unstructured protein domains. *Journal of Chromatography B*, 786(1–2), 247–254.

- Karbe, L. (1987). *Hot brines and the deep sea environment*. AJ Edwards, and SM Head (Oxford: Pergamon Press).
- Kato, D., & Oishi, M. (2014). Ultrasensitive detection of DNA and RNA based on enzyme-free click chemical ligation chain reaction on dispersed gold nanoparticles. *ACS Nano*, 8(10), 9988–9997.
- Kelch, B. A., & Agard, D. A. (2007). Mesophile versus thermophile: Insights into the structural mechanisms of kinetic stability. *Journal of Molecular Biology*, 370(4), 784–795.
- Kelley, L., & Sternberg, M. (2015). PHYRE2 Protein Fold Recognition Server. *Nature Protocols*, 10, 845–858.
- Keppetipola, N., & Shuman, S. (2005). Characterization of a Thermophilic ATP-Dependent DNA Ligase from the Euryarchaeon *Pyrococcus horikoshii*. *Journal of Bacteriology*, 187(20), 6902–6908. <https://doi.org/10.1128/JB.187.20.6902-6908.2005>
- Kirkwood, J., Hargreaves, D., O’Keefe, S., & Wilson, J. (2015). Using isoelectric point to determine the pH for initial protein crystallization trials. *Bioinformatics*, 31(9), 1444–1451.
- Korhonen, L., & Lindholm, D. (2004). The ubiquitin proteasome system in synaptic and axonal degeneration: A new twist to an old cycle. *The Journal of Cell Biology*, 165(1), 27–30.
- Korycka-Machala, M., Rychta, E., Brzostek, A., Sayer, H. R., Rumijowska-Galewicz, A., Bowater, R. P., & Dziadek, J. (2007). Evaluation of NAD<sup>+</sup>-dependent DNA ligase of mycobacteria as a potential target for antibiotics. *Antimicrobial Agents and Chemotherapy*, 51(8), 2888–2897.
- Kovacs, D., Kalmar, E., Torok, Z., & Tompa, P. (2008). Chaperone activity of ERD10 and ERD14, two disordered stress-related plant proteins. *Plant Physiology*, 147(1), 381–390.

- Kumar, Sandeep, Tsai, C.-J., & Nussinov, R. (2000). Factors enhancing protein thermostability. *Protein Engineering*, 13(3), 179–191.
- Kumar, Sudhir, Stecher, G., Li, M., Knyaz, C., & Tamura, K. (2018). MEGA X: Molecular evolutionary genetics analysis across computing platforms. *Molecular Biology and Evolution*, 35(6), 1547–1549.
- Lahiri, S. D., Gu, R.-F., Gao, N., Karantzeni, I., Walkup, G. K., & Mills, S. D. (2012). Structure guided understanding of NAD<sup>+</sup> recognition in bacterial DNA ligases. *ACS Chemical Biology*, 7(3), 571–580.
- Landegren, U. (1993). What's new: Ligation-based DNA diagnostics. *Bioessays*, 15(11), 761–765.
- Lee, J. Y., Chang, C., Song, H. K., Moon, J., Yang, J. K., Kim, H.-K., Kwon, S.-T., & Suh, S. W. (2000). Crystal structure of NAD<sup>+</sup>-dependent DNA ligase: Modular architecture and functional implications. *The EMBO Journal*, 19(5), 1119–1129.
- Lehman, I. R. (1974). DNA ligase: Structure, mechanism, and function. *Science*, 186(4166), 790–797.
- Li, D., Luo, R., Liu, C.-M., Leung, C.-M., Ting, H.-F., Sadakane, K., Yamashita, H., & Lam, T.-W. (2016). MEGAHIT v1.0: A fast and scalable metagenome assembler driven by advanced methodologies and community practices. *Methods*, 102, 3–11.
- Li, J., Chu, X., Liu, Y., Jiang, J.-H., He, Z., Zhang, Z., Shen, G., & Yu, R.-Q. (2005). A colorimetric method for point mutation detection using high-fidelity DNA ligase. *Nucleic Acids Research*, 33(19), e168–e168.

- Li, L.-Z., Xie, T.-H., Li, H.-J., Qing, C., Zhang, G.-M., & Sun, M.-S. (2007). Enhancing the thermostability of *Escherichia coli* L-asparaginase II by substitution with pro in predicted hydrogen-bonded turn structures. *Enzyme and Microbial Technology*, *41*(4), 523–527.
- Li, Z., Kessler, W., van den Heuvel, J., & Rinas, U. (2011). Simple defined autoinduction medium for high-level recombinant protein production using T7-based *Escherichia coli* expression systems. *Applied Microbiology and Biotechnology*, *91*(4), 1203.
- Lim, J.-H., Choi, J., Han, S.-J., Kim, S., Hwang, H.-Z., Jin, D.-K., Ahn, B.-Y., & Han, Y. (2001). Molecular cloning and characterization of thermostable DNA ligase from *Aquifex pyrophilus*, a hyperthermophilic bacterium. *Extremophiles*, *5*(3), 161–168.
- Linding, R., Jensen, L. J., Diella, F., Bork, P., Gibson, T. J., & Russell, R. B. (2003). Protein disorder prediction: Implications for structural proteomics. *Structure*, *11*(11), 1453–1459.
- Linding, R., Russell, R. B., Neduva, V., & Gibson, T. J. (2003). GlobPlot: Exploring protein sequences for globularity and disorder. *Nucleic Acids Research*, *31*(13), 3701–3708.
- Lohman, G. J., Bauer, R. J., Nichols, N. M., Mazzola, L., Bybee, J., Rivizzigno, D., Cantin, E., & Evans Jr, T. C. (2016). A high-throughput assay for the comprehensive profiling of DNA ligase fidelity. *Nucleic Acids Research*, *44*(2), e14–e14.
- Lorenz, P., & Eck, J. (2005). Metagenomics and industrial applications. *Nature Reviews Microbiology*, *3*(6), 510.
- Ludwig, W., Wallner, G., Tesch, A., & Klink, F. (1991). A novel eubacterial phylum: Comparative nucleotide sequence analysis of a *tuf*-gene of *Flexistipes sinusarabici*. *FEMS Microbiology Letters*, *78*(2–3), 139–143.

- Malash, M. N., Hussein, N. A., Muawia, S., Nasr, M. I., & Siam, R. (2020). An optimized protocol for high yield expression and purification of an extremophilic protein. *Protein Expression and Purification*, 169, 105585.
- Martinez, M. A., Woodcroft, B. J., Espinoza, J. C. I., Zayed, A. A., Singleton, C. M., Boyd, J. A., Li, Y.-F., Purvine, S., Maughan, H., & Hodgkins, S. B. (2019). Discovery and ecogenomic context of a global *Caldiserica*-related phylum active in thawing permafrost, *Candidatus Cryoserica* phylum nov., *Ca. Cryoserica* class nov., *Ca. Cryosericales* ord. Nov., *Ca. Cryoseriaceae* fam. Nov., comprising the four species *Cryosericum septentrionale* gen. Nov. Sp. Nov., *Ca. C. hinesii* sp. Nov., *Ca. C. odellii* sp. Nov., *Ca. C. terrychapinii* sp. Nov. *Systematic and Applied Microbiology*, 42(1), 54–66.
- Matsutani, M., Hirakawa, H., Nishikura, M., Soemphol, W., Ali, I. A. I., Yakushi, T., & Matsushita, K. (2011). Increased number of Arginine-based salt bridges contributes to the thermotolerance of thermotolerant acetic acid bacteria, *Acetobacter tropicalis* SKU1100. *Biochemical and Biophysical Research Communications*, 409(1), 120–124.
- Meuzelaar, H., Vreede, J., & Woutersen, S. (2016). Influence of Glu/Arg, Asp/Arg, and Glu/Lys salt bridges on  $\alpha$ -helical stability and folding kinetics. *Biophysical Journal*, 110(11), 2328–2341.
- Michaelis, W., Jenisch, A., & Richnow, H. H. (1990). Hydrothermal petroleum generation in Red Sea sediments from the Kebrit and Shaban Deeps. *Applied Geochemistry*, 5(1–2), 103–114.
- Miller, A. R., Densmore, C. D., Degens, E. T., Hathaway, J. C., Manheim, F. T., McFarlin, P. F., Pocklington, R., & Jokela, A. (1966). Hot brines and recent iron deposits in deeps of the Red Sea. *Geochimica et Cosmochimica Acta*, 30(3), 341–359.

- Miller, E. S., Kutter, E., Mosig, G., Arisaka, F., Kunisawa, T., & Rüger, W. (2003). Bacteriophage T4 genome. *Microbiol. Mol. Biol. Rev.*, 67(1), 86–156.
- Monnin, C., & Ramboz, C. (1996). The anhydrite saturation index of the ponded brines and sediment pore waters of the Red Sea deeps. *Chemical Geology*, 127(1–3), 141–159.
- Mortazavi, M., & Hosseinkhani, S. (2011). Design of thermostable luciferases through arginine saturation in solvent-exposed loops. *Protein Engineering, Design & Selection*, 24(12), 893–903.
- Nandakumar, J., Nair, P. A., & Shuman, S. (2007). Last stop on the road to repair: Structure of *E. coli* DNA ligase bound to nicked DNA-adenylate. *Molecular Cell*, 26(2), 257–271.
- Nishida, H., Kiyonari, S., Ishino, Y., & Morikawa, K. (2006). The closed structure of an archaeal DNA ligase from *Pyrococcus furiosus*. *Journal of Molecular Biology*, 360(5), 956–967.
- Niu, B., Fu, L., Sun, S., & Li, W. (2010). Artificial and natural duplicates in pyrosequencing reads of metagenomic data. *BMC Bioinformatics*, 11(1), 187.
- Noguchi, H., Taniguchi, T., & Itoh, T. (2008). MetaGeneAnnotator: Detecting species-specific patterns of ribosomal binding site for precise gene prediction in anonymous prokaryotic and phage genomes. *DNA Research*, 15(6), 387–396.
- Odell, M., & Shuman, S. (1999). Footprinting of *Chlorella* virus DNA ligase bound at a nick in duplex DNA. *Journal of Biological Chemistry*, 274(20), 14032–14039.
- Pace, N. R. (2009). Mapping the tree of life: Progress and prospects. *Microbiol. Mol. Biol. Rev.*, 73(4), 565–576.
- Pautot, G., Guennoc, P., Coutelle, A., & Lyberis, N. (1984). Discovery of a large brine deep in the northern Red Sea. *Nature*, 310(5973), 133.

- Pergolizzi, G., Butt, J. N., Bowater, R. P., & Wagner, G. K. (2011). A novel fluorescent probe for NAD-consuming enzymes. *Chemical Communications*, 47(47), 12655–12657.
- Pergolizzi, G., Wagner, G. K., & Bowater, R. P. (2016). Biochemical and structural characterization of DNA ligases from bacteria and archaea. *Bioscience Reports*, 36(5), e00391.
- Piel, J. (2002). A polyketide synthase-peptide synthetase gene cluster from an uncultured bacterial symbiont of *Paederus* beetles. *Proceedings of the National Academy of Sciences*, 99(22), 14002–14007.
- Psifidi, A., Dovas, C., & Banos, G. (2011). Novel quantitative real-time LCR for the sensitive detection of SNP frequencies in pooled DNA: Method development, evaluation and application. *PloS One*, 6(1).
- Reysenbach, A.-L., Wickham, G. S., & Pace, N. R. (1994). Phylogenetic analysis of the hyperthermophilic pink filament community in Octopus Spring, Yellowstone National Park. *Applied and Environmental Microbiology*, 60(6), 2113–2119.
- Ross, D. A. (1972). Red Sea hot brine area: Revisited. *Science*, 175(4029), 1455–1457.
- Sælensminde, G., Halskau, Ø., & Jonassen, I. (2009). Amino acid contacts in proteins adapted to different temperatures: Hydrophobic interactions and surface charges play a key role. *Extremophiles*, 13(1), 11.
- Sakaguchi, M., Matsuzaki, M., Niimiya, K., Seino, J., Sugahara, Y., & Kawakita, M. (2007). Role of proline residues in conferring thermostability on aqualysin I. *The Journal of Biochemistry*, 141(2), 213–220.
- Sayed, A., Ghazy, M. A., Ferreira, A. J., Setubal, J. C., Chambergo, F. S., Ouf, A., Adel, M., Dawe, A. S., Archer, J. A., & Bajic, V. B. (2014). A novel mercuric reductase from the



- unique deep brine environment of Atlantis II in the Red Sea. *Journal of Biological Chemistry*, 289(3), 1675–1687.
- Schmieder, R., & Edwards, R. (2011). Quality control and preprocessing of metagenomic datasets. *Bioinformatics*, 27(6), 863–864.
- Scholten, J. C., Staffers, P., Garbe-Schdnberg, D., & Moammar, M. (2017). Hydrothermal mineralization in the Red Sea. In *Handbook of Marine Mineral Deposits* (pp. 369–395). Routledge.
- Scott, B. O., Lavesa-Curto, M., Bullard, D. R., Butt, J. N., & Bowater, R. P. (2006). Immobilized DNA hairpins for assay of sequential breaking and joining of DNA backbones. *Analytical Biochemistry*, 358(1), 90–98.
- Seitz, K. W., Dombrowski, N., Eme, L., Spang, A., Lombard, J., Sieber, J. R., Teske, A. P., Ettema, T. J. G., & Baker, B. J. (2019). *Asgard archaea capable of anaerobic hydrocarbon cycling. Nat Commun 10: 1822.*
- Sgaramella, V., Van de Sande, J. H., & Khorana, H. G. (1970). Studies on polynucleotides, CA novel joining reaction catalyzed by the T4-polynucleotide ligase. *Proceedings of the National Academy of Sciences*, 67(3), 1468–1475.
- Sharma, A. K., Ali, A., Gogna, R., Singh, A. K., & Pati, U. (2009). P53 amino-terminus region (1–125) stabilizes and restores heat denatured p53 wild phenotype. *PLoS One*, 4(10).
- Shen, W., Deng, H., & Gao, Z. (2012). Gold nanoparticle-enabled real-time ligation chain reaction for ultrasensitive detection of DNA. *Journal of the American Chemical Society*, 134(36), 14678–14681.
- Shuman, S. (2009). DNA ligases: Progress and prospects. *Journal of Biological Chemistry*, 284(26), 17365–17369.

- Shuman, S., & Lima, C. D. (2004). The polynucleotide ligase and RNA capping enzyme superfamily of covalent nucleotidyltransferases. *Current Opinion in Structural Biology*, 14(6), 757–764.
- Siam, R., Mustafa, G. A., Sharaf, H., Moustafa, A., Ramadan, A. R., Antunes, A., Bajic, V. B., Stingl, U., Marsis, N. G., & Coolen, M. J. (2012). Unique prokaryotic consortia in geochemically distinct sediments from Red Sea Atlantis II and Discovery Deep brine pools. *PloS One*, 7(8), e42872.
- Simon, C., & Daniel, R. (2009). Achievements and new knowledge unraveled by metagenomic approaches. *Applied Microbiology and Biotechnology*, 85(2), 265–276.
- Singh, J., Behal, A., Singla, N., Joshi, A., Birbian, N., Singh, S., Bali, V., & Batra, N. (2009). Metagenomics: Concept, methodology, ecological inference and recent advances. *Biotechnology Journal: Healthcare Nutrition Technology*, 4(4), 480–494.
- Smith, D. K., Radivojac, P., Obradovic, Z., Dunker, A. K., & Zhu, G. (2003). Improved amino acid flexibility parameters. *Protein Science*, 12(5), 1060–1072.
- Sokalingam, S., Raghunathan, G., Soundrarajan, N., & Lee, S.-G. (2012). A study on the effect of surface lysine to arginine mutagenesis on protein stability and structure using green fluorescent protein. *PloS One*, 7(7).
- Somero, G. N., Lockwood, B. L., & Tomanek, L. (2017). *Biochemical adaptation: Response to environmental challenges, from life's origins to the Anthropocene*. Sinauer Associates, Incorporated Publishers.
- Spang, A., Saw, J. H., Jørgensen, S. L., Zaremba-Niedzwiedzka, K., Martijn, J., Lind, A. E., van Eijk, R., Schleper, C., Guy, L., & Ettema, T. J. (2015). Complex archaea that bridge the gap between prokaryotes and eukaryotes. *Nature*, 521(7551), 173.

- Steele, H. L., Jaeger, K.-E., Daniel, R., & Streit, W. R. (2009). Advances in recovery of novel biocatalysts from metagenomes. *Journal of Molecular Microbiology and Biotechnology*, *16*(1–2), 25–37.
- Stejskalová, E., Horáková, P., Vacek, J., Bowater, R. P., & Fojta, M. (2014). Enzyme-linked electrochemical DNA ligation assay using magnetic beads. *Analytical and Bioanalytical Chemistry*, *406*(17), 4129–4136.
- Stewart, J., Kozlowski, P., Sowden, M., Messing, E., & Smith, H. C. (1998). A quantitative assay for assessing allelic proportions by iterative gap ligation. *Nucleic Acids Research*, *26*(4), 961–966.
- Strickler, S. S., Gribenko, A. V., Gribenko, A. V., Keiffer, T. R., Tomlinson, J., Reihle, T., Loladze, V. V., & Makhatadze, G. I. (2006). Protein stability and surface electrostatics: A charged relationship. *Biochemistry*, *45*(9), 2761–2766.
- Strub, C., Alies, C., Lougarre, A., Ladurantie, C., Czaplicki, J., & Fournier, D. (2004). Mutation of exposed hydrophobic amino acids to arginine to increase protein stability. *BMC Biochemistry*, *5*(1), 1–6.
- SUZUKI, Y. (1989). A general principle of increasing protein thermostability. *Proceedings of the Japan Academy, Series B*, *65*(6), 146–148.
- Tantos, A., Friedrich, P., & Tompa, P. (2009). Cold stability of intrinsically disordered proteins. *FEBS Letters*, *583*(2), 465–469.
- Tatusova, T., Ciufu, S., Fedorov, B., O'Neill, K., & Tolstoy, I. (2014). RefSeq microbial genomes database: New representation and annotation strategy. *Nucleic Acids Research*, *42*(D1), D553–D559.

- Thompson, M. J., & Eisenberg, D. (1999). Transproteomic evidence of a loop-deletion mechanism for enhancing protein thermostability. *Journal of Molecular Biology*, 290(2), 595–604.
- Timson, D. J., Singleton, M. R., & Wigley, D. B. (2000). DNA ligases in the repair and replication of DNA. *Mutation Research/DNA Repair*, 460(3–4), 301–318.
- Tompa, P. (2002). Intrinsically unstructured proteins. *Trends in Biochemical Sciences*, 27(10), 527–533.
- Tripathi, R. P., Pandey, J., Kukshal, V., Ajay, A., Mishra, M., Dube, D., Chopra, D., Dwivedi, R., Chaturvedi, V., & Ramachandran, R. (2011). Synthesis, in silico screening and bioevaluation of dispiro-cycloalkanones as antitubercular and mycobacterial NAD<sup>+</sup>-dependent DNA ligase inhibitors. *MedChemComm*, 2(5), 378–384.
- Trüper, H. G. (1969). Bacterial sulfate reduction in the Red Sea hot brines. In *Hot brines and recent heavy metal deposits in the Red Sea* (pp. 263–271). Springer.
- Tully, B. J., Graham, E. D., & Heidelberg, J. F. (2018). The reconstruction of 2,631 draft metagenome-assembled genomes from the global oceans. *Scientific Data*, 5, 170203.
- Tully, B. J., Wheat, C. G., Glazer, B. T., & Huber, J. A. (2018). A dynamic microbial community with high functional redundancy inhabits the cold, oxic subsurface aquifer. *The ISME Journal*, 12(1), 1–16.
- Uchiyama, T., & Miyazaki, K. (2009). Functional metagenomics for enzyme discovery: Challenges to efficient screening. *Current Opinion in Biotechnology*, 20(6), 616–622.
- Uria, A. R., Fawzya, Y. N., & Chasanah, E. (2005). Novel molecular methods for discovery and engineering of biocatalysts from uncultured marine microorganisms. *J. Coast. Develop*, 8(2), 49–74.

- Uversky, V. N., Gillespie, J. R., & Fink, A. L. (2000). Why are “natively unfolded” proteins unstructured under physiologic conditions? *Proteins: Structure, Function, and Bioinformatics*, *41*(3), 415–427.
- Vacek, J., Cahova, K., Palecek, E., Bullard, D. R., Lavesa-Curto, M., Bowater, R. P., & Fojta, M. (2008). Label-free electrochemical monitoring of DNA ligase activity. *Analytical Chemistry*, *80*(19), 7609–7613.
- Van der Linden, M. G., & de Farias, S. T. (2006). Correlation between codon usage and thermostability. *Extremophiles*, *10*(5), 479–481.
- Van Dijk, E. L., Jaszczyszyn, Y., & Thermes, C. (2014). Library preparation methods for next-generation sequencing: Tone down the bias. *Experimental Cell Research*, *322*(1), 12–20.
- Wang, Y., Cao, H., Zhang, G., Bougouffa, S., Lee, O. O., Al-Suwailem, A., & Qian, P.-Y. (2013). Autotrophic microbe metagenomes and metabolic pathways differentiate adjacent Red Sea brine pools. *Scientific Reports*, *3*, 1748.
- Watanabe, K., Chishiro, K., Kitamura, K., & Suzuki, Y. (1991). Proline residues responsible for thermostability occur with high frequency in the loop regions of an extremely thermostable oligo-1, 6-glucosidase from *Bacillus thermoglucosidasius* KP1006. *Journal of Biological Chemistry*, *266*(36), 24287–24294.
- Watson, S. W., & Waterbury, J. B. (1969). The sterile hot brines of the Red Sea. In *Hot brines and recent heavy metal deposits in the Red Sea* (pp. 272–281). Springer.
- Weiss, B., & Rich, C. C. (1967). ARD so N, Enzymatic breakage and joining of deoxyribonucleic acid, I. Repair of single-strand breaks in DNA by an enzyme system from *Escherichia coli* infected with T4 bacteriophage. *Proc. Vall. Acad. Sci. US*, *57*, 1021–1028.

- Wiedmann, M., Wilson, W. J., Czajka, J., Luo, J., Barany, F., & Batt, C. A. (1994). Ligase chain reaction (LCR)-overview and applications. *PCR Methods Appl*, 3(4), S51–64.
- Wilkinson, A., Day, J., & Bowater, R. (2001). Bacterial DNA ligases. *Molecular Microbiology*, 40(6), 1241–1248.
- Williamson, A., Hjerde, E., & Kahlke, T. (2016). Analysis of the distribution and evolution of the ATP-dependent DNA ligases of bacteria delineates a distinct phylogenetic group ‘Lig E.’ *Molecular Microbiology*, 99(2), 274–290.
- Wilson, R. H., Morton, S. K., Deiderick, H., Gerth, M. L., Paul, H. A., Gerber, I., Patel, A., Ellington, A. D., Hunicke-Smith, S. P., & Patrick, W. M. (2013). Engineered DNA ligases with improved activities in vitro. *Protein Engineering, Design & Selection*, 26(7), 471–478.
- Winckler, G., Aeschbach-Hertig, W., Kipfer, R., Botz, R., Rübél, A. P., Bayer, R., & Stoffers, P. (2001). Constraints on origin and evolution of Red Sea brines from helium and argon isotopes. *Earth and Planetary Science Letters*, 184(3–4), 671–683.
- Wolcott, M. J. (1992). Advances in nucleic acid-based detection methods. *Clinical Microbiology Reviews*, 5(4), 370–386.
- Wright, P. E., & Dyson, H. J. (1999). Intrinsically unstructured proteins: Re-assessing the protein structure-function paradigm. *Journal of Molecular Biology*, 293(2), 321–331.
- Wu, D. Y., & Wallace, R. B. (1989). The ligation amplification reaction (LAR)—Amplification of specific DNA sequences using sequential rounds of template-dependent ligation. *Genomics*, 4(4), 560–569.
- Xu, M., Fujita, D., & Hanagata, N. (2009). Perspectives and challenges of emerging single-molecule DNA sequencing technologies. *Small*, 5(23), 2638–2649.

- Yadav, N., Khanam, T., Shukla, A., Rai, N., Hajela, K., & Ramachandran, R. (2015). Tricyclic dihydrobenzoxazepine and tetracyclic indole derivatives can specifically target bacterial DNA ligases and can distinguish them from human DNA ligase I. *Organic & Biomolecular Chemistry*, 13(19), 5475–5487.
- Yang, W., Yang, Y., Zhang, L., Xu, H., Guo, X., Yang, X., Dong, B., & Cao, Y. (2017). Improved thermostability of an acidic xylanase from *Aspergillus sulphureus* by combined disulphide bridge introduction and proline residue substitution. *Scientific Reports*, 7(1), 1–9.
- Yokota, K., Satou, K., & Ohki, S. (2006). Comparative analysis of protein thermostability: Differences in amino acid content and substitution at the surfaces and in the core regions of thermophilic and mesophilic proteins. *Science and Technology of Advanced Materials*, 7(3), 255.
- Yount, B., Denison, M. R., Weiss, S. R., & Baric, R. S. (2002). Systematic assembly of a full-length infectious cDNA of mouse hepatitis virus strain A59. *Journal of Virology*, 76(21), 11065–11078.
- Yu, H., & Huang, H. (2014). Engineering proteins for thermostability through rigidifying flexible sites. *Biotechnology Advances*, 32(2), 308–315.
- Yu, H., Zhao, Y., Guo, C., Gan, Y., & Huang, H. (2015). The role of proline substitutions within flexible regions on thermostability of luciferase. *Biochimica et Biophysica Acta (BBA)-Proteins and Proteomics*, 1854(1), 65–72.
- Zhang, C.-Y., Wu, Z.-Q., Yin, D.-C., Zhou, B.-R., Guo, Y.-Z., Lu, H.-M., Zhou, R.-B., & Shang, P. (2013). A strategy for selecting the pH of protein solutions to enhance crystallization.



*Acta Crystallographica Section F: Structural Biology and Crystallization Communications*, 69(7), 821–826.

Zhao, A., Gray, F. C., & MacNeill, S. A. (2006). ATP- and NAD<sup>+</sup>-dependent DNA ligases share an essential function in the halophilic archaeon *Haloferax volcanii*. *Molecular Microbiology*, 59(3), 743–752.

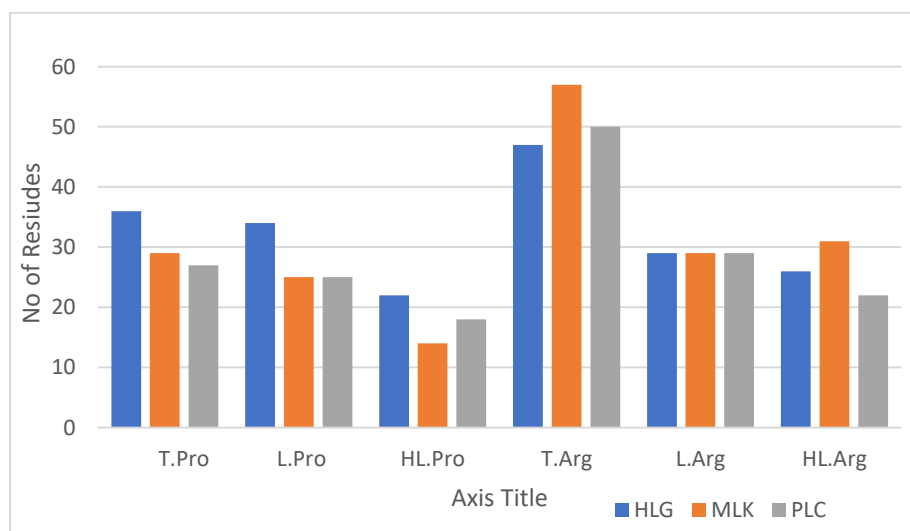
Zhou, C., Xue, Y., & Ma, Y. (2010). Enhancing the thermostability of  $\alpha$ -glucosidase from *Thermoanaerobacter tengcongensis* MB4 by single proline substitution. *Journal of Bioscience and Bioengineering*, 110(1), 12–17.

Zhou, Q., Su, X., Jing, G., & Ning, K. (2014). Meta-QC-Chain: Comprehensive and fast quality control method for metagenomic data. *Genomics, Proteomics & Bioinformatics*, 12(1), 52–56.

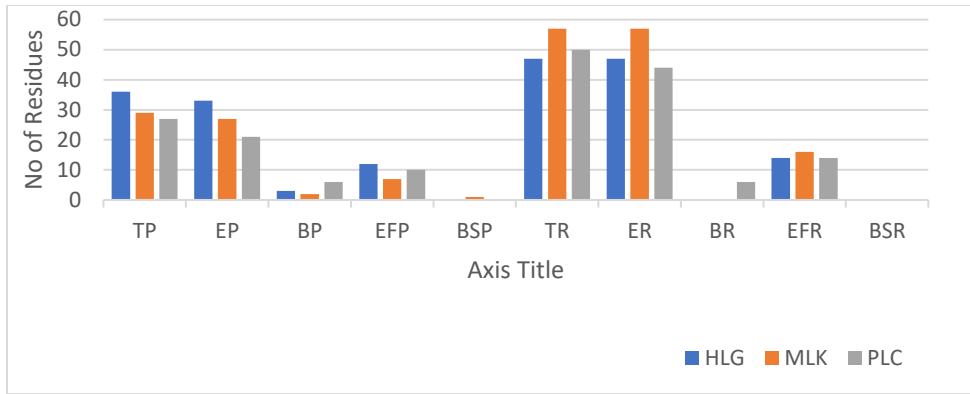
Supplementary information

**Supplementary Figures: Proline and Arginine analyses at the loops, hot loops, exposed and buried regions**

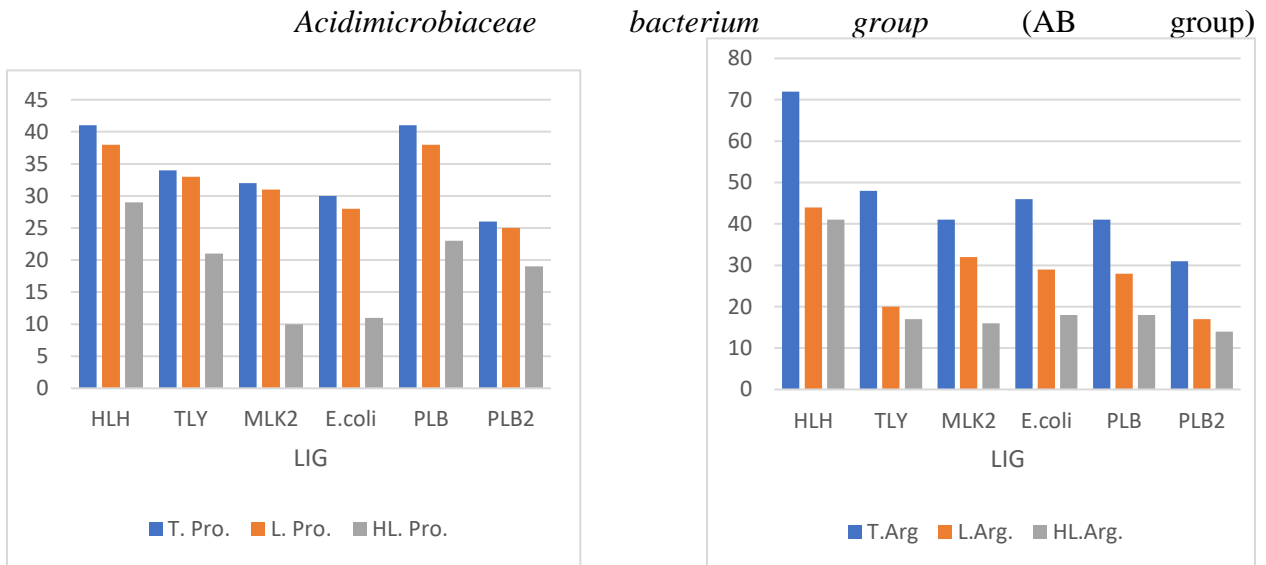
*Candidatus Marinimicrobia bacterium* group (CBa group)



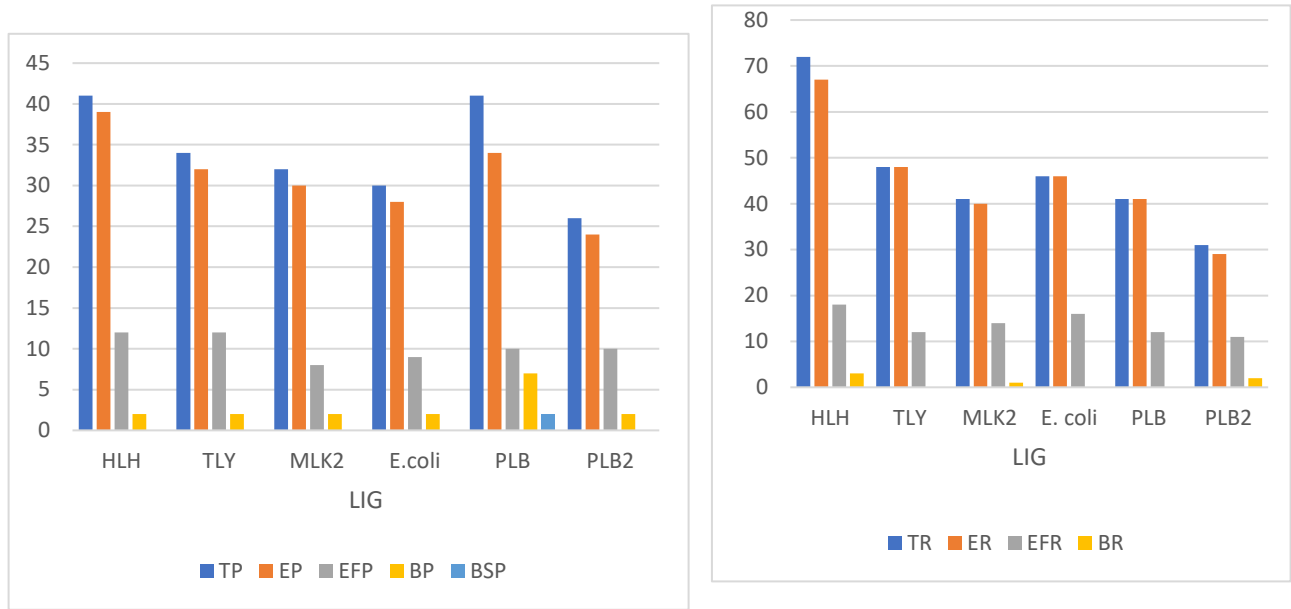
**Fig.1 Proline and Arginine residues analysis in the primary structures, loop/coils and hot loops regions of ligase sequences belonging to the *Candidatus marinimicrobia bacterium***



**Fig 2. Proline and Arginine residues analysis in the primary structures, exposed and buried regions of ligase sequences belonging to the *Candidatus marinimicrobia bacterium***

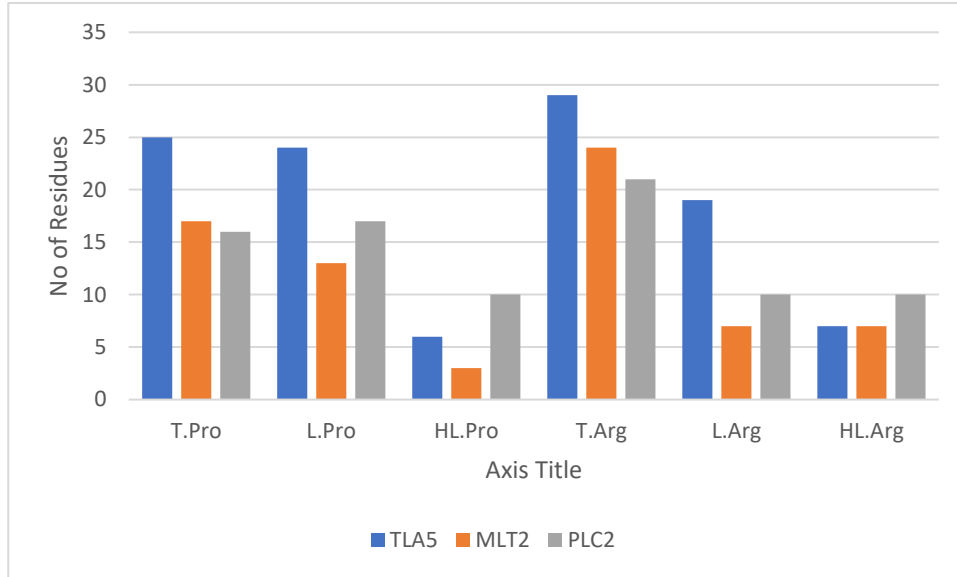


**Fig 3 Proline and Arginine residues analysis in the primary structures, loop/coils and hot loops regions of ligase sequences belonging to the *Acidimicrobiaceae* bacterium species**

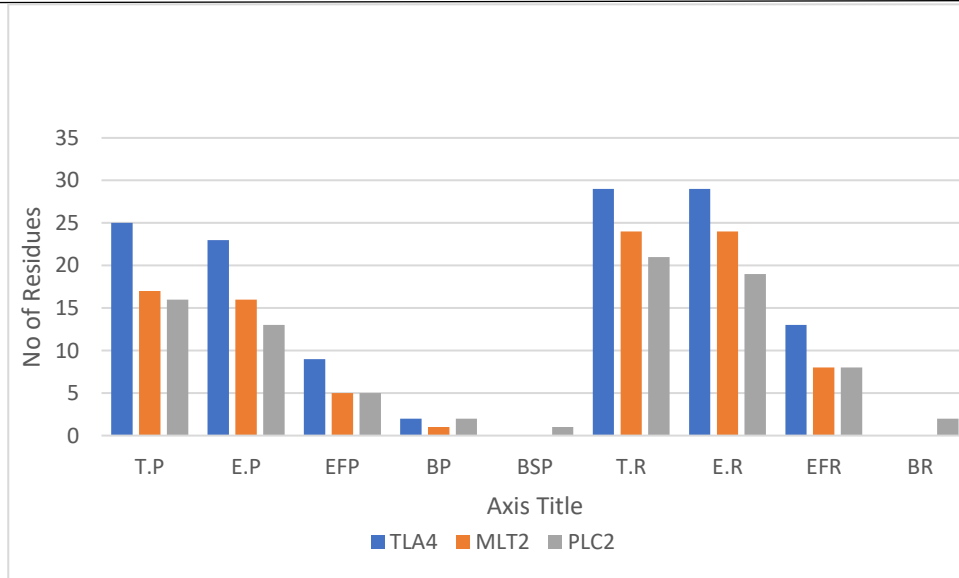


**Fig.4 Proline and Arginine residues analysis in the primary structures, exposed and buried regions of ligase sequences belonging to the *Acidimicrobiaceae* bacterium**

*Moraxallaceae bacterium group (MB group)*

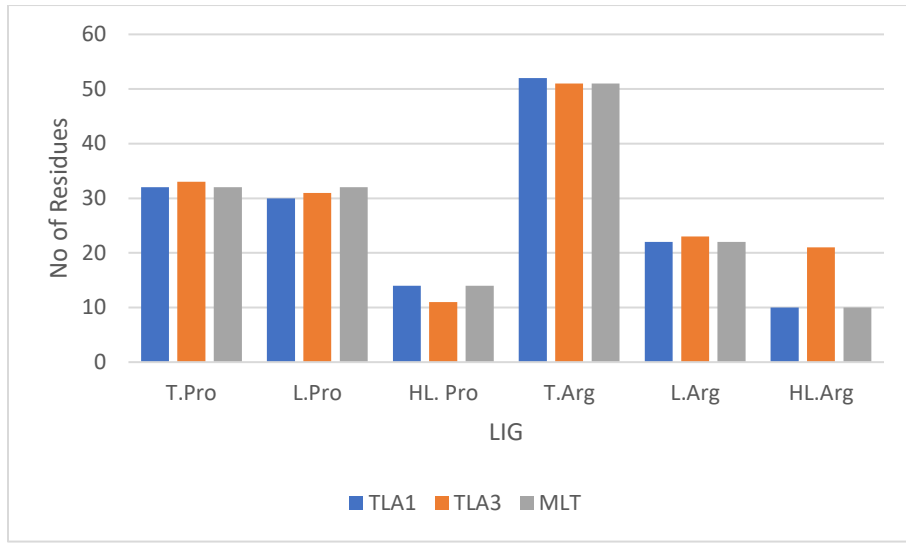


**Fig.5 Proline and Arginine residues analysis in the primary structures, loop/coils and hot loops regions of ligase sequences belonging to the *Moraxallaceae bacterium***

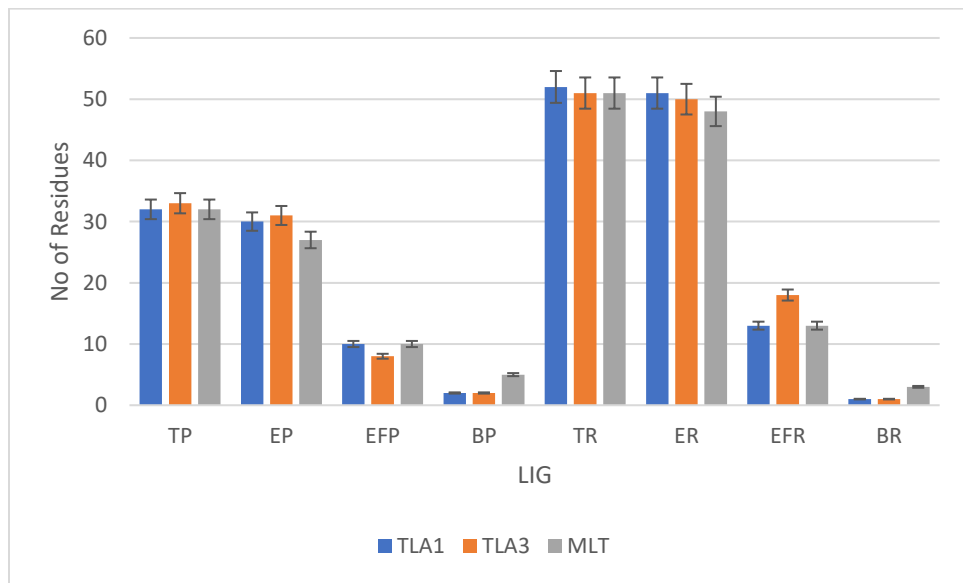


**Fig.6 Proline and Arginine residues analysis in the primary structures, loop/coils and hot loops regions of ligase sequences belonging to the *Moraxallaceae bacterium***

*Phyllobacterium myrsinacearum* group

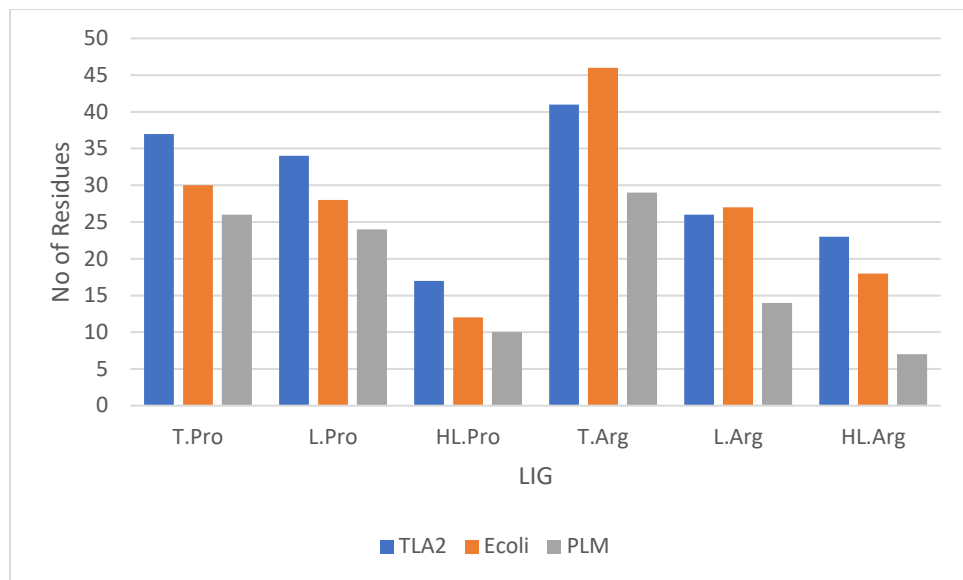


**Fig.7 Proline and Arginine residues analysis in the primary structures, loop/coils and hot loops regions of ligase sequences belonging to the *Phyllobacterium myrsinacearum***

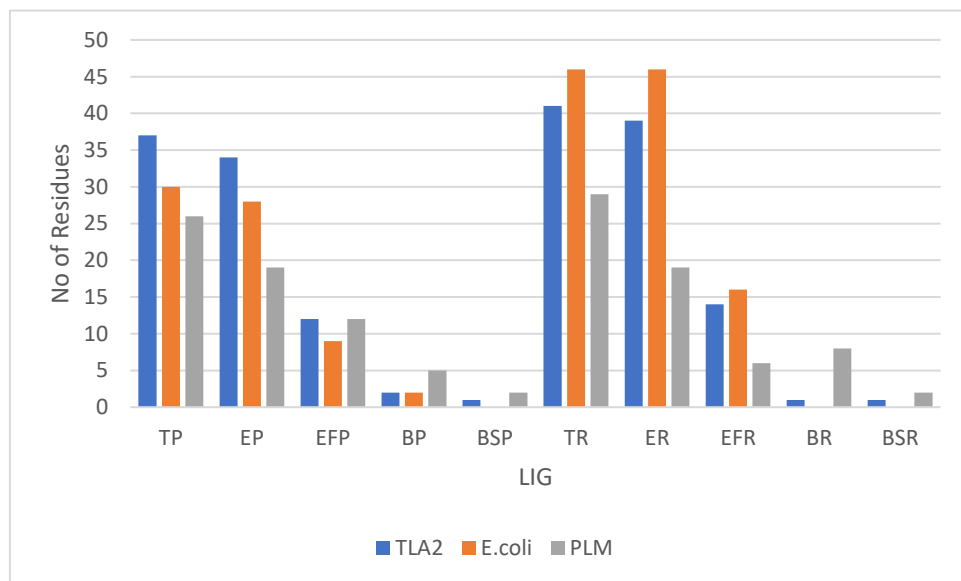


**Fig.8 Proline and Arginine residues analysis in the primary structures, loop/coils and hot loops regions of ligase sequences belonging to the *Phyllobacterium myrsinacearum* species**

***Rhizobiales bacterium group (RB group)***



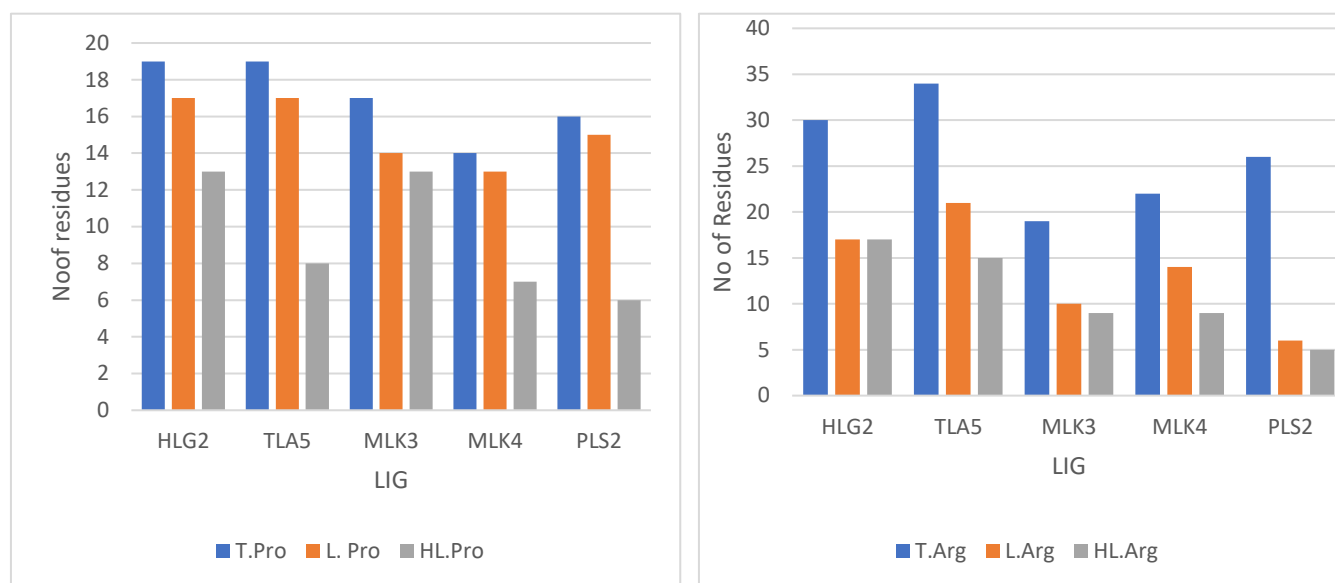
**Fig.9) Proline and Arginine residues analysis in the primary structures, loop/coils and hot loops regions of ligase sequences belonging to the *Rhizobiales bacterium species***



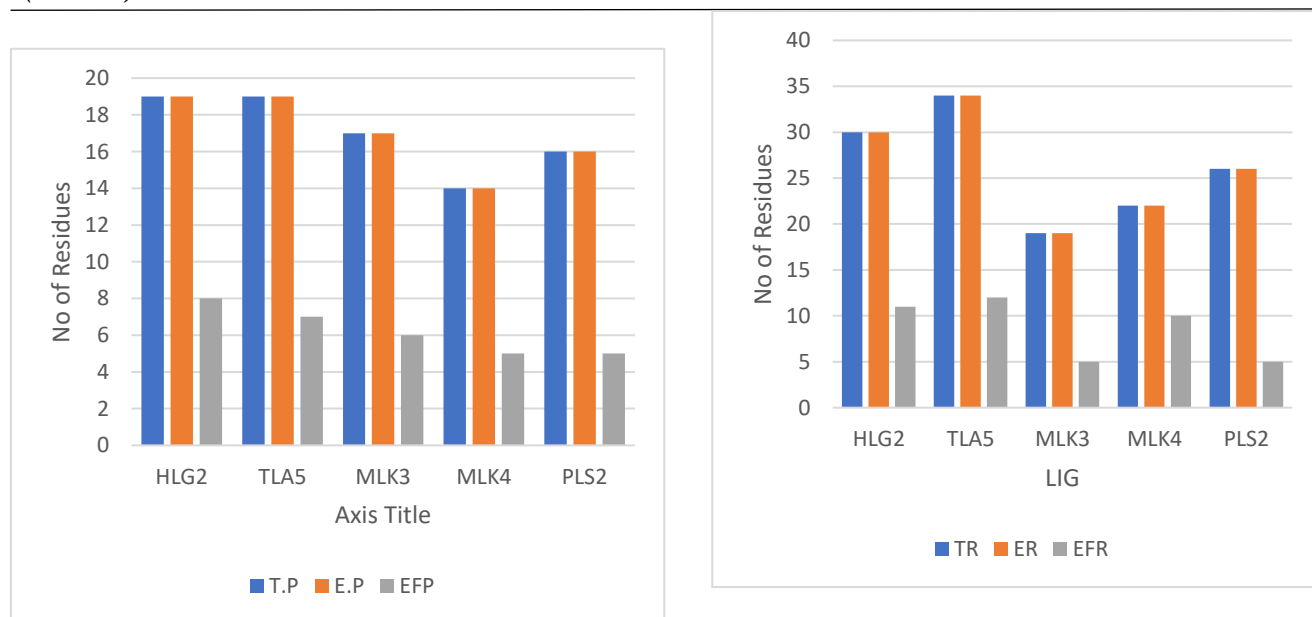
**Fig.10) Proline and Arginine residues analysis in the primary structures, exposed and buried regions of ligase sequences belonging to the *Rhizobiales bacterium species***



*Candidatus marinimicrobia bacterium(Partial)(CBb group)*



**Fig.11) Proline and Arginine residues analysis in the primary structures, loop/coils and hot loops regions of ligase sequences belonging to the *Candidatus marinimicrobia bacterium (Partial)***



**Fig.12) Proline and Arginine residues analysis in the primary structures, exposed and buried regions of ligase sequences belonging to the *Candidatus marinimicrobia bacterium(partial)***

Supplementary figures (1-12): Proline and Arginine residues analysis in the primary and secondary structures of ligase sequences grouped under various selected species group. **TP/T.Pro** represents the total proline, **TR/T.Arg** represents the total arginine residues, **L.Pro** represents the Proline residues in the loops/coils, **HL.Pro** represents Proline residues at the Hot-loops regions, **LArg** represents the Arginine residues at the loops/coils, **HL.Arg** represents the Arginine residues at the Hot-loops regions, **EP** represents the Exposed Proline residues, **ER** represents the Exposed Arginine residues, **BP** represents the Buried Proline residues, **EFP** represents the Exposed functional proline, **EFR** represents the Exposed functional Arginine residues, **BSP** represents the Exposed Buried Proline residues, **BSR** represents the Buried Structural Arginine Residues. The loop and hot-loop regions analysis was done in the DisEMBL server while the Functionality region analysis (exposed& buried) was done on ConSurf server. The error-bars were estimated by percentages.