1 CHAPTER 1: INTRODUCTION

1.1 PURPOSE OF STUDY

The study aims to analyse the evolution of the metalloenzyme alcohol dehydrogenase of the eukaryote *Mus musculus* (mouse) through database similarity searching and multiple sequence alignment. The focus would be on identifying the conserved site across a group of bacterial sequences and to identify their phylogenetic tree.

1.2 MOLECULAR EVOLUTION

Molecular evolution is the study of the evolutionary patterns and processes of nucleic acids and proteins, as well as the use of these molecules in studies of phylogenetics, population genetics, biogeography, and other areas of research at the DNA, RNA and protein levels of the biological organization (MolecularEvolution.org, 2012). The field of molecular biology emerged as a scientific field from the convergence of work by researchers from molecular biology, evolutionary biology and population genetics on understanding a common problem that is the structure and function of the gene particularly on enzyme function, species divergence and the origin of noncoding DNA (Molecular Biology, 2010 & Molecular Evolution Wikipedia, 2012).

1.3 METALLOENZYME

Metalloenzyme is an enzyme that contains a bound metal ion as part of its structure. The metal may be required for enzymic activity, either participating directly in catalysis or stabilising the active conformation of the protein (Knox, 1955).

1.4 CONSERVED SITE

Conserved site is where the sequences of amino acids in DNA or RNA remain similar or identical across multiple species (Conserved Sequence NCBI, 2003). Protein residues that are identified to be critical for structure and function are predicted to be conserved throughout evolution (Schueler-Furman, 2003). An enzyme's active site is highly conserved and participates in the catalytic activity of the protein.

1.5 OBJECTIVES OF STUDY

There are three main objectives in this study:

1. To determine the secondary structure of the 1E3I protein
2. To identify the Alcohol Dehydrogenase's active site
3. To determine the phylogenetic relationship of Alcohol Dehydrogenase among different species

2 CHAPTER 2: LITERATURE REVIEW

2.1 INTRODUCTION TO ALCOHOL DEHYDROGENASE

Alcohol dehydrogenase (ADH) is a part of the general classes of enzymes named oxireductases. The ADH is a kind of zinc metalloenzyme made up of a group of seven dehydrogenase enzymes occurring in many organisms. The enzyme contains unvaryingly large amounts of zinc strongly bound to the protein where the zinc metal acts as a functional component of the molecule in its enzymatic activity (Vallee & Hoch, 1955).

The ADH constitutes a natural family varying from plant ADH, animal ADH, and FADH (a class III ADH known as the glutathione-dependent formaldehyde hydrogenase) (Persson, Hendlund & Jornvall, 2008). ADH catalyses the reversible reduction of ketones and aldehydes to their corresponding alcohols (Adhikary, Ganguli & Jyoti, 2010). A coenzyme such as nicotinamide adenine dinucleotide (NAD^+) or its phosphate ($NADP^+$) as the electron receptor is required to facilitate the interconversion of the alcohols by way of reduction (Musa & Philips, 2011).

Human ADH is divided into 5 classes encoded by 7 genes as shown in Table 2.1. The human ADH is a dimeric enzyme where for each of the classes, the protein products are similar in amino acid sequence and structures but varies in preferred substrates (Edenberg, 2000). The majority of ethanol metabolism is performed by the

ADHs in Class 1 consisting of ADH1A (alpha), ADH1B (beta), and ADH1C (gamma), the ADHs in Class 2 which is ADH4; and the ADH in Class 4 which is ADH7. The variety of Class 1 ADHs is found in the mucosa of the stomach, liver, lung, kidney, and lower digestive tract. The Class 2 ADHs can be found in the liver and the Class 4 ADHs are found in the mucosa of the upper digestive tract and stomach. ADH activity is inhibited by the high levels of the products of ADH-mediated ethanol oxidation, acetaldehyde and NADH (Green & Stoler, 2007).

2.2 MECHANISM OF ALCOHOL DEHYDROGENASE

In humans and animals, ADH together with Aldehyde Dehydrogenase (ALDH) breaks down alcohols which would otherwise be toxic to the body (Kops, 2009). ADH in some bacteria and yeast catalyses the opposite reaction as part of fermentation where it converts acetaldehyde to ethanol (Bamforth & Kanauchi, 2004). ADH to some extent participates in the cell defence towards exogenous alcohols and aldehydes (Hoog, Hedberg, Stromberg & Svensson, 2001).

Through fermentation, bacteria are capable of producing energy under anaerobic conditions. In a reductive reaction mediated by bacterial ADH, ethanol is derived from acetaldehyde as the end product of alcoholic fermentation (Neale et al, 1986). The reaction catalysed by microbial ADH can run in the opposite direction when there is an excess of ethanol where acetaldehyde will be the end product (Jokelainen, Roine, Vaananen, Farkkila & Salaspuro, 1994).

Table 2.1: The Human Alcohol Dehydrogenase (ADH) Genes and Proteins.

Official Gene Name	**Old Name**	**NCBI Reference Sequence**	**Protein**	**Class**
ADH1A	*ADH1*	NM_000667	α	I
ADH1B	*ADH2*	NM_000668	β	I
ADH1C	*ADH3*	NM_000669	γ	I
ADH4	*ADH4*	NM_000670	π	II
ADH5	*ADH5*	NM_000671	χ	III
ADH6	*ADH6*	NM_000672	ADH6	V
ADH7	*ADH7*	NM_000673	σ	IV

2.3 THE ALCOHOL DEHYDROGENESE ACTIVE SITE

Each chain of the alcohol dehydrogenase active site consists of one NAD^+ and two Zn^{2+} binding sites. An essential aspect of ADH's catalysis is the electrostatic stabilization of the alcohol's oxygen by the zinc atom. This makes the proton on the alcohol more acidic. Each Zinc ion is ligated directly between the side chains of Cys 46, His 67, Cys 174 and a water molecule which is hydrogen bonded to Thr 48. (Ryan, Matt & Martin, 1999)

There are two clefts between the two binding sites where the zinc is located. One of the cleft binds NAD^+ and the other binds the ethanol. NAD^+ is the coenzyme for ADH and plays a significant role in the conversion of ethanol. The NAD^+ is bound by multiple residues off the Rossman fold which is a series of beta-alpha-beta motif (Lesk, 1995). Some of the amino acid residues that bind NAD^+ are Arg 47, Try 178, Val 203, Gly 210, Asp 223, Asn 225, Asn 242, Pro 243, Val 268 and Gly 292.

One NAD^+ molecule is used to convert ethanol to acetaldehyde by proton transfer. ADH catalyses the oxidation of alcohols by reducing NAD with a hydride. The usage of a zinc ion to electrostatically stabilize the alcohol oxygen increases the acidity of the alcohol's proton. During this process, His 51 is activated by general base catalysis where the histidine accepts a proton from the NAD. The NAD then acquires a proton from another general base catalysis which is the Thr 48.

During hydrogen transfer, two hydrogens are stripped off the ethanol by zinc. These proton transfers prepare the negatively charged threonine for accepting a proton from the alcohol of the actual substrate. Since this oxidation is rigorous, there is a hydride transfer to the NAD^+ in its traditional hydride accepting region at the same time (Prasad, 2011).

The NAD^+ that binds at residues 293-298 causes a 10° rotation and gets the catalytic domain to shift closer to the coenzyme binding domain and closes the active site cleft (Ramaswamy, Park & Plapp, 1999). The two active sites are in clefts between the coenzyme binding core and the catalytic domains. The ethanol binds to the hydrophobic core lined by nine amino acids surrounding the substrate. After binding NAD^+, the 10° rotation makes the protein go from its open form to the closed form. The coenzyme binding domains forming the centre of the dimer molecule retained their conformation and orientation within the molecule and hence narrows the cleft. The narrowing of cleft brings the substrate binding site closer and excludes water from the active site which is vital for the activity of ADH (Eklund & Branden, 1979).

The process described is in essence the transfer of hydride to the NAD^+ and the oxidation of an alcohol to an aldehyde. The orientation of the amino acid proton acceptors and donors during catalysis and the coordinate of the zinc ion relative to the substrate are important in influencing the electrostatic potential and transition state stabilization in the active site (Baker *et al.*, 2009).

3 CHAPTER 3: MATERIALS AND METHODS

3.1 MATERIALS

The first step in this research was that an alcohol dehydrogenase class II protein structure with the PDB ID 1E3I was identified and extracted from the Protein Data Bank (RCSB PDB).

Then, the analysis of the protein structure was done using the necessary hardware and software listed below:

3.1.1 Hardware

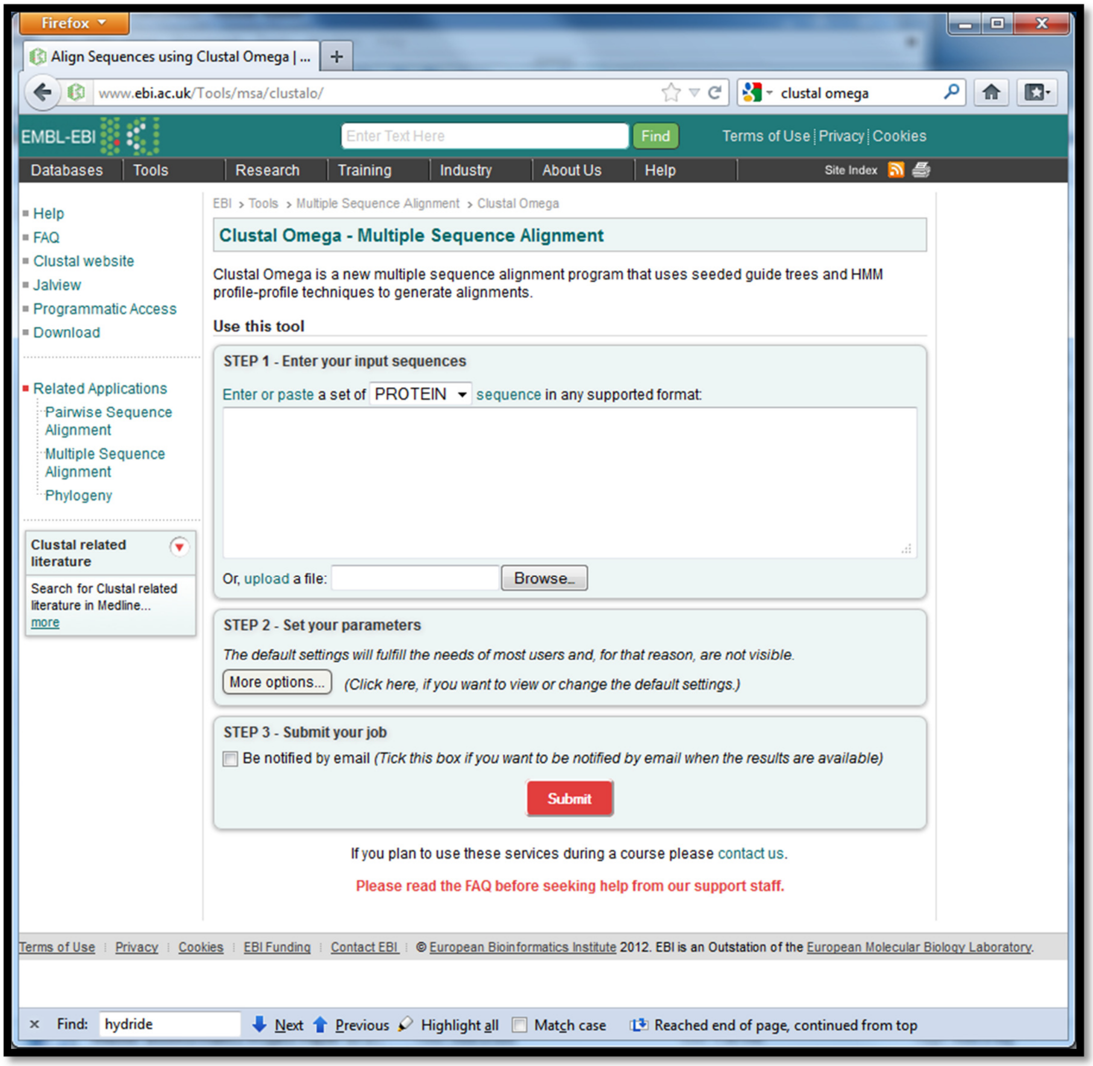
Two main computers were used

1. Intel® Core™ i5-2400 CPU 3.10GHz, 4.00 GB RAM
2. Intel Core 2 Duo 1.86GHz, 4.00 GB RAM

3.1.2 Software

Table 3.1: The Software being Used.

Software	Description
RCSB PDB Ligand Explorer	Visualize details of the chemical components of structures. The Ligand Explorer was designed for rapid inspections of protein-ligand interactions, such as hydrophilic interactions, hydrophobic interactions and the interactions with ordered H_2O molecules.
Clustal Omega, Clustal W and Clustal X	Clustal series of programs for multiple sequence alignment for alignment of proteins, DNA and RNA. Available both as web server based (Clustal Omega and Clustal W) and downloadable (Clustal X with Graphical User Interface and Clustal W with command line). The Clustal Omega screen shot can be seen in Figure 3.1.
PhyloDraw	PhyloDraw is a tool for producing phylogenetic trees. PhyloDraw supports a variety of multialignment programs (Clustal-W, Phylip format, Dialign2, and pairwise distance matrix) and visualizes different types of tree diagrams
Rasmol Version 2.7.5	Molecular Visualization tool for proteins, DNA and macromolecules. Originally developed by Roger Sayle in the early 90s.



Clustal Omega - Multiple Sequence Alignment

Clustal Omega is a new multiple sequence alignment program that uses seeded guide trees and HMM profile-profile techniques to generate alignments.

Use this tool

STEP 1 - Enter your input sequences

Enter or paste a set of PROTEIN sequence in any supported format:

Or, upload a file:

STEP 2 - Set your parameters

The default settings will fulfill the needs of most users and, for that reason, are not visible.

More options... *(Click here, if you want to view or change the default settings.)*

STEP 3 - Submit your job

□ Be notified by email *(Tick this box if you want to be notified by email when the results are available)*

If you plan to use these services during a course please contact us.

Please read the FAQ before seeking help from our support staff.

Search for Clustal related literature in Medline...

Terms of Use | Privacy | Cookies | EBI Funding | Contact EBI | © European Bioinformatics Institute 2012. EBI is an Outstation of the European Molecular Biology Laboratory.

Find: hydride

Reached end of page, continued from top

Figure 3.1: The Clustal Omega Multiple Sequence Alignment Web Based Program on EMBL-EBI.

3.2 METHODS

Figure 3.2 shows the flowchart of the methodology used in this research. There were three main parts involved which can be divided into:

Task 1: Determine Phylogenetic Relationship of Alcohol Dehydrogenase

Task 2: Identify Consensus Sequences and Trace Analysis

Task 3: Protein Mapping

In brief, the methods involved were:

1. Identify protein of a eukaryote containing ADH
2. Retrieve PDB file of 1E3I from RCSB Protein Data Bank
3. Identify and select similar bacterial sequences from BLAST
4. Do multiple sequence alignment using Clustal
5. Construct phylogenetic tree using PhyloDraw
6. Define the partitions by drawing vertical lines across the tree
7. Determine groups to analyse and to drop out of analysis
8. Construct consensus sequences
9. Perform evolutionary trace thus identify conserved, neutral and class-specific residues (if any)
10. Determine amino acid residues within 5A of each of the organic ligands NAD and CXF

11. Compare trace analysis findings with the active site residues in PDB, Rasmol analysis and literature readings
12. Map the trace analysis residues identified onto the protein structure

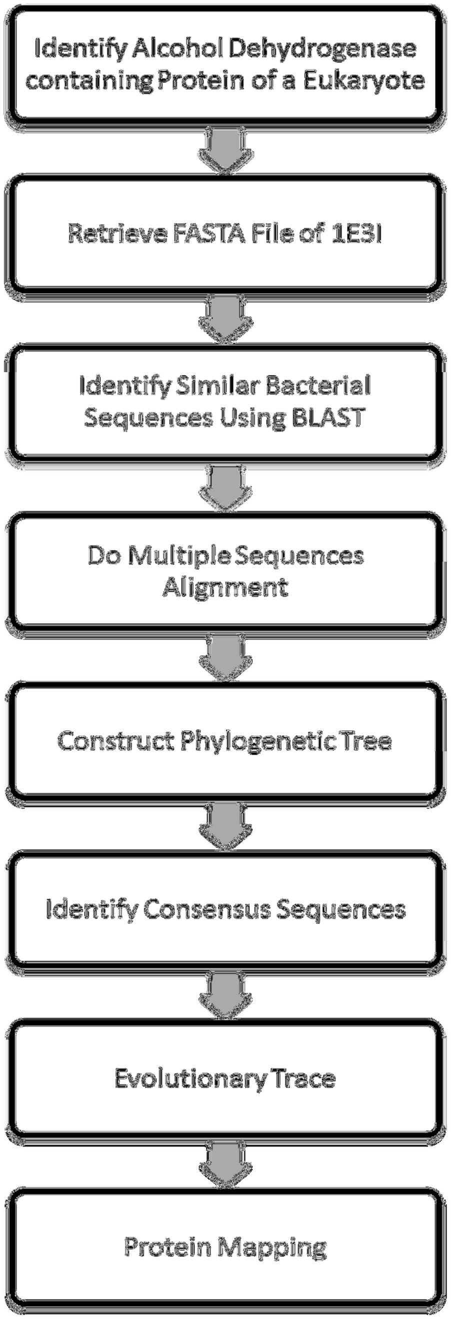

Identify Alcohol Dehydrogenase containing Protein of a Eukaryote
Retrieve FASTA File of 1E3I
Identify Similar Bacterial Sequences Using BLAST
Do Multiple Sequences Alignment
Construct Phylogenetic Tree
Identify Consensus Sequences
Evolutionary Trace
Protein Mapping

Figure 3.2: The Flowchart of the Methodology.

3.3 TASK 1: DETERMINING THE PHYLOGENETIC RELATIONSHIP OF ALCOHOL DEHYDROGENASE

3.3.1 Identification of Alcohol Dehydrogenase Containing Protein

The protein of a eukaryote containing the Alcohol Dehydrogenase enzyme which is a type of zinc metalloenzyme was identified. The 1E3I protein from the *Mus musculus* (mouse) was selected as the protein to be studied. The 1E3I PDB file was retrieved from the RCSB Protein Data Bank at:

"http://www.pdb.org/pdb/explore/explore.do?structureId=1e3i"

3.3.2 Verifying 1E3I FASTA Sequence from PDB with UniProt

The FASTA file retrieved from PDB containing the sequence with 376 Amino Acid Residues was compared with the sequence in UniProt. The UniProt identifier for 1E3I is Q9QYY9 and was retrieved from Swiss-Prot which is in the UniProt Knowledgebase from the link below:

"http://www.uniprot.org/uniprot/Q9QYY9"

3.3.3 BLAST Search

The 1E3I sequence was then inserted into the Basic Local Alignment Search Tool for protein database using a protein query (protein blast). The Standard Protein BLAST using the blastp algorithm was used to identify similar bacteria sequences from the database of Non-redundance protein sequences (nr). A screenshot of the BLAST query page can be seen in Figure 3.3. The blast link is as below:

"http://blast.ncbi.nlm.nih.gov/Blast.cgi"

48 out of the 100 bacteria sequences were chosen for analysis. The ADH bacteria sequences chosen were based on the sequences with the high percentage of similarity compared to the *Mus musculus* 1E3I sequence.

3.3.4 Collecting FASTA Format Sequences of Bacteria

The FASTA format sequences of the identified bacteria were taken from the BLAST results for further analysis.

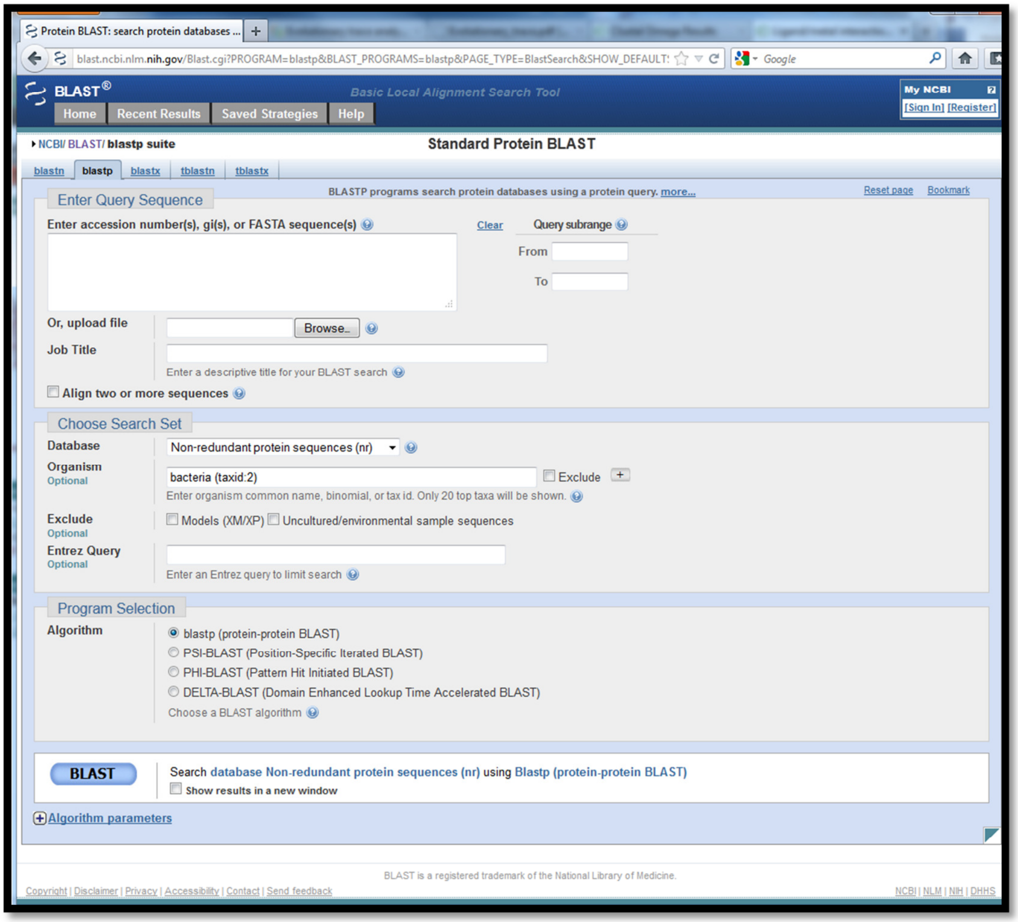
3.3.5 Multiple Sequence Alignment

The ClustalX program which is the graphical version of the ClustalW command line version was used to for the multiple sequences alignment (MSA) of all 48 bacteria sequences and both chains of the *Mus musculus* ADH sequences identified resulting in a total of 50 sequences. Complete alignments of all 50 sequences were done.

3.3.6 Construction of Phylogenetic Tree

The Trees menu in the ClustalX program was used to generate a Phylip Tree file using Neighbor-Joining Method (N-J Tree).

The .ph file generated from the multiple sequence alignment is used to construct the phylogenetic tree on PhyloDraw. The rectangular cladogram tree was constructed using cladistics method and from a pairwise distance matrix using the Neighbor Joining clustering method.



Protein BLAST: search protein databases ...

blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastp&BLAST_PROGRAMS=blastp&PAGE_TYPE=BlastSearch&SHOW_DEFAULT:

BLAST® Basic Local Alignment Search Tool

Home | Recent Results | Saved Strategies | Help

My NCBI [Sign In] [Register]

NCBI/ BLAST/ **blastp suite**

Standard Protein BLAST

blastn | **blastp** | blastx | tblastn | tblastx

BLASTP programs search protein databases using a protein query. more... Reset page Bookmark

Enter Query Sequence

Enter accession number(s), gi(s), or FASTA sequence(s) Clear

Query subrange
From
To

Or, upload file Browse...

Job Title

Enter a descriptive title for your BLAST search

□ Align two or more sequences

Choose Search Set

Database: Non-redundant protein sequences (nr)

Organism (Optional): bacteria (taxid:2) □ Exclude +

Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown.

Exclude (Optional): □ Models (XM/XP) □ Uncultured/environmental sample sequences

Entrez Query (Optional):

Enter an Entrez query to limit search

Program Selection

Algorithm
● blastp (protein-protein BLAST)
○ PSI-BLAST (Position-Specific Iterated BLAST)
○ PHI-BLAST (Pattern Hit Initiated BLAST)
○ DELTA-BLAST (Domain Enhanced Lookup Time Accelerated BLAST)
Choose a BLAST algorithm

BLAST Search database Non-redundant protein sequences (nr) using Blastp (protein-protein BLAST)
□ Show results in a new window

+ Algorithm parameters

BLAST is a registered trademark of the National Library of Medicine.

Copyright | Disclaimer | Privacy | Accessibility | Contact | Send feedback   NCBI | NLM | NIH | DHHS

Figure 3.3: Standard Protein BLAST Suite.

3.4 TASK 2: IDENTIFYING CONSENSUS SEQUENCES AND TRACE ANALYSIS

3.4.1 Partitioning the Cladogram and Species Grouping

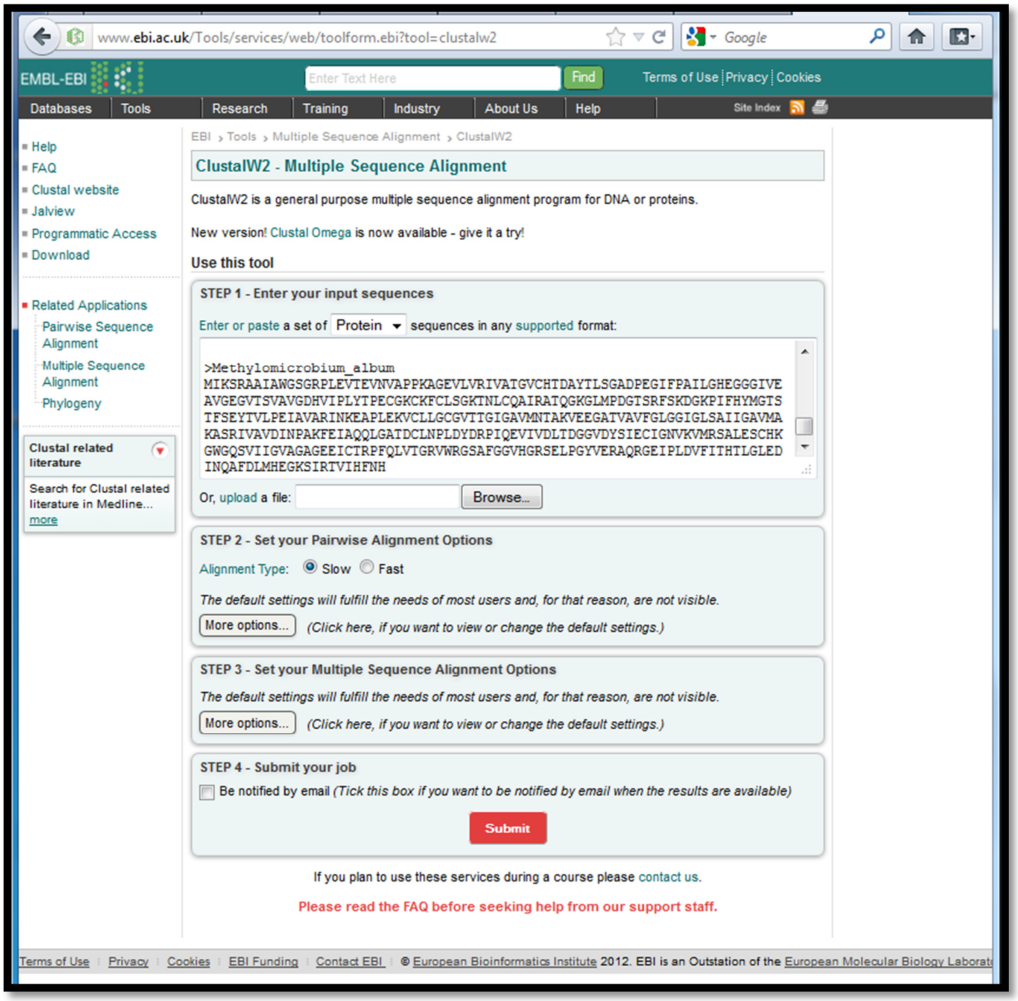
The Rectangular Cladogram constructed from Task 1 was used for further analysis in Task 2. Vertical lines were made on the tree to partition the sequences. A partition of the entire family is generated by grouping together sequences that branch off from a node. 5 vertical lines were made across the dendogram and were identified as P0, P1, P2, P3 and P4.

3.4.2 Single Group Multiple Sequences Alignment

Each partition was analysed separately beginning from P0. The groups from partition P0 consisting of Group 1 and Group 2 were each separately aligned using the Web Based Clustal. The ClustalW2 Multiple Sequence Alignment program on the EMBL-EBI server was used to align each of the groups generated from the partitioning from the link below and the interface can be seen as in Figure 3.4:

"http://www.clustal.org/clustal2/"

The multiple sequence alignment was repeated for each of the group identified up to partition P4.



www.ebi.ac.uk/Tools/services/web/toolform.ebi?tool=clustalw2
Google
EMBL-EBI
Enter Text Here
Find
Terms of Use | Privacy | Cookies
Databases
Tools
Research
Training
Industry
About Us
Help
Site Index
Help
FAQ
Clustal website
Jalview
Programmatic Access
Download
Related Applications
Pairwise Sequence Alignment
Multiple Sequence Alignment
Phylogeny
Clustal related literature
Search for Clustal related literature in Medline...
more
EBI › Tools › Multiple Sequence Alignment › ClustalW2

ClustalW2 - Multiple Sequence Alignment

ClustalW2 is a general purpose multiple sequence alignment program for DNA or proteins.

New version! Clustal Omega is now available - give it a try!

Use this tool

STEP 1 - Enter your input sequences

Enter or paste a set of Protein sequences in any supported format:

```
>Methylomicrobium_album
MIKSRAAIAWGSGRPLEVTEVNVAPPKAGEVLVRIVATGVCHTDAYTLSGADPEGIFPAILGHEGGGIVE
AVGEGVTSVAVGDHVIPLYTPECGKCKFCLSGKTNLCQAIRATQGKGLMPDGTSRFSKDGKPIFHYMGTS
TFSEYTVLPEIAVARINKEAPLEKVCLLGCGVTTGIGAVMNTAKVEEGATVAVFGLGGIGLSAIIGAVMA
KASRIVAVDINPAKFEIAQQLGATDCLNPLDYDRPIQEVIVDLTDGGVDYSIECIGNVKVMRSALESCHK
GWGQSVIIGVAGAGEEICTRPFQLVTGRVWRGSAFGGVHGRSELPGYVERAQRGEIPLDVFITHTLGLED
INQAFDLMHEGKSIRTVIHFNH
```

Or, upload a file: Browse...

STEP 2 - Set your Pairwise Alignment Options

Alignment Type: ● Slow ○ Fast

The default settings will fulfill the needs of most users and, for that reason, are not visible.

More options... *(Click here, if you want to view or change the default settings.)*

STEP 3 - Set your Multiple Sequence Alignment Options

The default settings will fulfill the needs of most users and, for that reason, are not visible.

More options... *(Click here, if you want to view or change the default settings.)*

STEP 4 - Submit your job

☐ Be notified by email *(Tick this box if you want to be notified by email when the results are available)*

Submit

If you plan to use these services during a course please contact us.

Please read the FAQ before seeking help from our support staff.

Terms of Use | Privacy | Cookies | EBI Funding | Contact EBI | © European Bioinformatics Institute 2012. EBI is an Outstation of the European Molecular Biology Laborat

Figure 3.4: The ClustalW2 Multiple Sequence Alignment Program on EMBL-EBI Server.

3.4.1 Identification of Consensus Sequences and Evolutionary Trace

A consensus sequence was then constructed for each group by labelling positions that are invariant in the multiple sequence alignment by its amino acid. The variable positions were left blank.

The invariant residues were then compared between the consensus sequences generated from each group. The conserved residues in the entire family or class-specific residues which were residues that changed between the groups or neutral residues which were residues that failed to be invariant in at least one group were identified. The conserved residues were being translated into the trace record exactly as they were. The class-specific residues if any were identified with an X. The neutral residues were identified with blank positions.

3.5 TASK 3: PROTEIN MAPPING

3.5.1 Secondary Structure of 1E3I

The Rasmol software is used in this task and a screen shot of the program is as Figure 3.5. As both the ADH chains were identified to be identical, the analysis was done using only chain A. The secondary structure of the 1E3I protein was viewed at consecutive 90° rotation about the X-axis consisting of 4 different views at 0°, 90°, 180° and 270°.

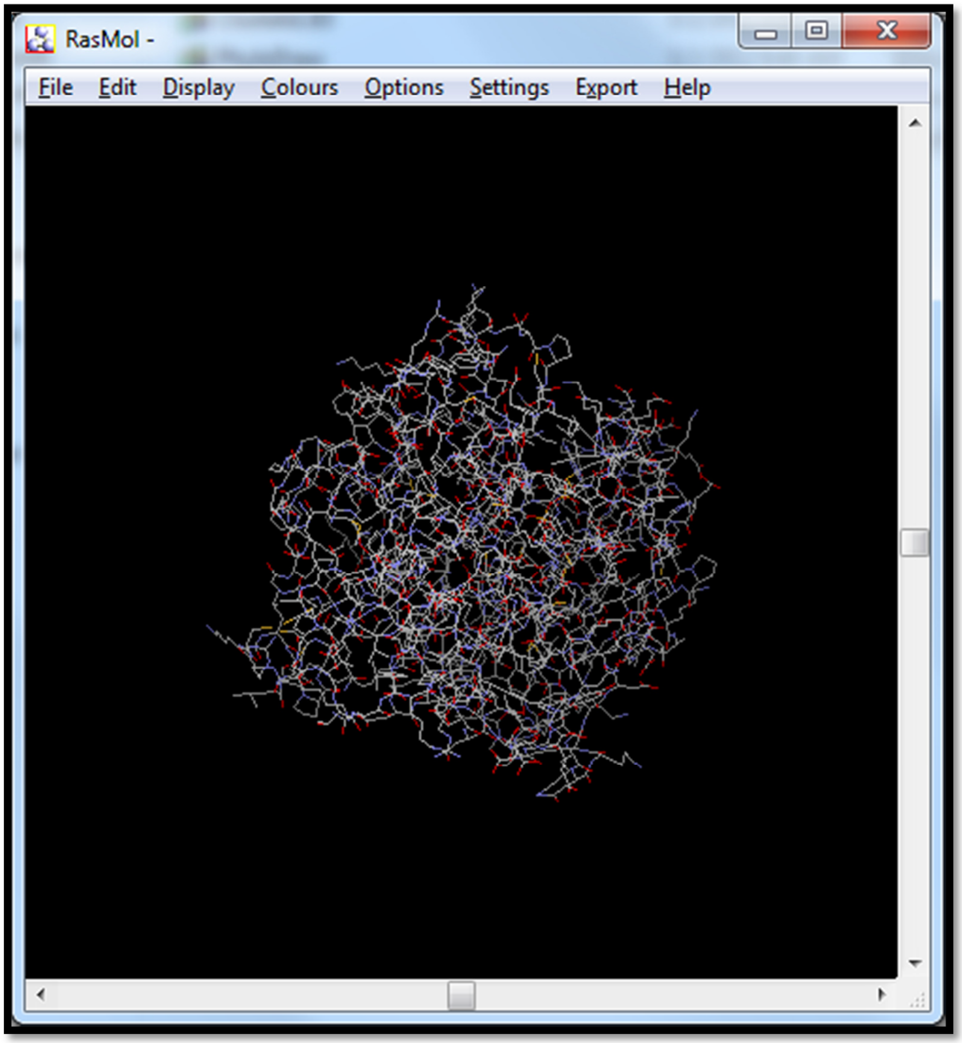


Figure 3.5: The Rasmol Molecular Visualization Tool.

3.5.2 The Zinc Atom, NAD Ligand and CXF Ligand with Surrounding Residues

Rasmol was used to identify the amino acid residues within 5 Armstrong of the organic ligands NAD and CXF using the following commands:

```
restrict *:a
centre *:a
Menu Display> Cartoons

select NAD
wireframe 100
select within (5.0, NAD)
Menu Display> Ball & Stick

select CXF
wireframe 100
select within (5.0, CXF)
Menu Display> Ball & Stick

select ZN
color gold
Menu Display> Spacefill
```

3.5.3 Comparison of Trace Analysis, Rasmol and Active Sites Identified from PDB SUM

The alcohol dehydrogenase active sites were identified from PDB SUM.

The amino acid residues identified from the Trace Analysis, Rasmol and Active Sites Identified from PDB SUM were compared in a table for Partition P0 and Partition P4.

Amino acid residues which exist in Rasmol analysis and conserved in the trace analysis were highlighted in green. The amino acid residues which were PDB SUM active site residues and also conserved residues from the evolutionary trace were highlighted in blue. When the amino acid residues exist in all three columns the amino acid residues were highlighted in yellow.

3.5.4 Mapping of Conserved Residues on the Protein Structure

The conserved amino acid residues from the Trace Analysis were mapped onto the protein structure in Rasmol. The conserved residues were identified in green.

4 CHAPTER 4: RESULTS

4.1 TASK 1: DETERMINING THE PHYLOGENETIC RELATIONSHIP OF ALCOHOL DEHYDROGENASE

4.1.1 Identification of Alcohol Dehydrogenase Containing Protein

The Alcohol dehydrogenase containing protein of a *Mus musculus* has been identified from the PDB. The secondary structure viewed using Jmol is as Figure 4.1.

4.1.2 Verifying 1E3I FASTA Sequence from PDB with UniProt

The FASTA sequence with the PDB ID 1E3I is as shown in Figure 4.2. The sequence is compared with the sequence in UniProt with the identifier Q9QYY9. The sequence has a length of 376 amino acid residues with two identical chains A and B.

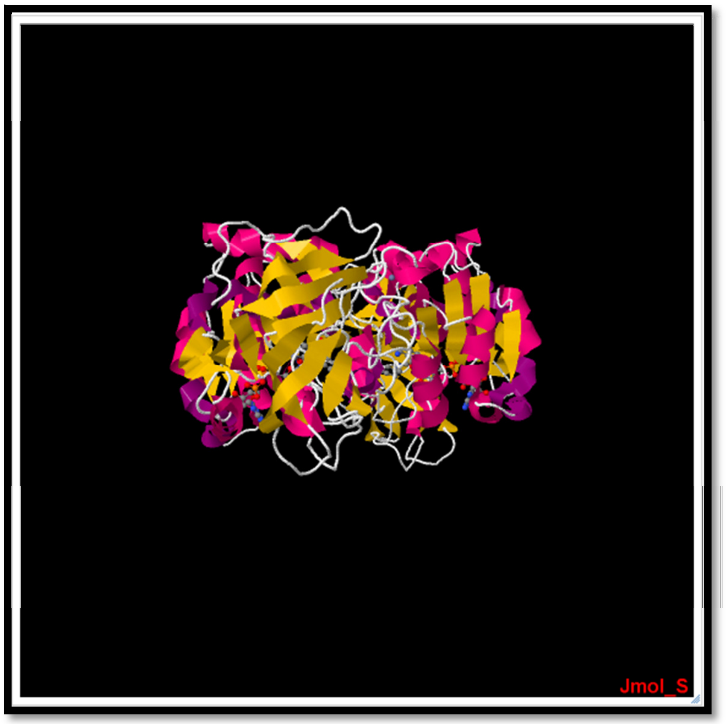


Figure 4.1: Secondary Structure of 1E3I Viewed using Jmol. Pink represents the Alpha Helix and Yellow represents the Beta Sheets.

```
>1E3I:A|PDBID|CHAIN|SEQUENCE

GTQGKVIKCKAAIAWKTGSPLCIEEIEVSPPKACEVRIQVIATCVCPTDINATDPKKKALFPVVLGHECAGIVESVGPGV

TNFKPGDKVIPFFAPQCKRCKLCLSPLTNLCGKLRNFKYPTIDQELMEDRTSRFTCKGRSIYHFMGVSSFSQYTVVSEAN

LARVDDEANLERVCLIGCGFSSGYGAAINTAKVTPGSTCAVFGLGCVGLSAIIGCKIAGASRIIAIDINGEKFPKAKALG

ATDCLNPRELDKPVQDVITELTAGGVDYSLDCAGTAQTLKAAVDCTVLGWGSCTVVGAKVDEMTIPTVDVILGRSINGTF

FGGWKSVDSVPNLVSDYKNKKFDLDLLVTHALPFESINDAIDLMKEGKSIRTILTF


>1E3I:B|PDBID|CHAIN|SEQUENCE

GTQGKVIKCKAAIAWKTGSPLCIEEIEVSPPKACEVRIQVIATCVCPTDINATDPKKKALFPVVLGHECAGIVESVGPGV

TNFKPGDKVIPFFAPQCKRCKLCLSPLTNLCGKLRNFKYPTIDQELMEDRTSRFTCKGRSIYHFMGVSSFSQYTVVSEAN

LARVDDEANLERVCLIGCGFSSGYGAAINTAKVTPGSTCAVFGLGCVGLSAIIGCKIAGASRIIAIDINGEKFPKAKALG

ATDCLNPRELDKPVQDVITELTAGGVDYSLDCAGTAQTLKAAVDCTVLGWGSCTVVGAKVDEMTIPTVDVILGRSINGTF

FGGWKSVDSVPNLVSDYKNKKFDLDLLVTHALPFESINDAIDLMKEGKSIRTILTF
```

Figure 4.2: FASTA Sequence of 1E3I.

4.1.3 BLAST Search

100 bacteria sequences from the BLAST search are being generated from the 1E3I query sequence. Out of the 100 bacteria sequences, 48 bacteria sequences are chosen originating from 48 different bacteria species with the highest similarity to the query sequence. The remaining sequences of bacteria from the same species already in the list are eliminated. All 48 are ADH bacteria sequences with similarity values ranging between 48% and 52% compared to the 1E3I query sequence. The organisms and NCBI reference for the sequences are as in Table 4.1.

4.1.4 Collecting FASTA Format Sequences of Bacteria

The FASTA format sequences of all 48 bacteria are obtained for further analysis. The FASTA format sequence identifier following the ">" (more than) symbol for each of the sequence is replaced with only the organism name removing all other description in the original FASTA format sequence. The two parts of the scientific name of the bacteria are connected with an "_" (underscore) symbol. Some sample FASTA format sequences are shown as in Figure 4.3.

Table 4.1: BLAST Results of Selected Similar Bacterial Sequences. The Mus musculus Query Sequence is Highlighted in Purple at the Top of the Table.

Accession	**Organism**	**Max ident**
CAB57455.1	*Mus musculus*	100%
YP_003147226.1	*Kangiella koreensis*	52%
NP_900410.1	*Chromobacterium violaceum*	51%
ZP_08518499.1	*Aeromonas caviae*	51%
ZP_09161412.1	*Marinobacter manganoxydans*	51%
ZP_09505218.1	*Alteromonas sp. S89*	51%
YP_959517.1	*Marinobacter aquaeolei*	51%
YP_005428879.1	*Marinobacter hydrocarbonoclasticus*	51%
ZP_00956464.1	*Sulfitobacter sp. EE-36*	51%
YP_001675317.1	*Shewanella halifaxensis*	51%
YP_001502873.1	*Shewanella pealeana*	51%
ZP_01893942.1	*Marinobacter algicola*	51%
ZP_09508887.1	*Marinobacterium stanieri*	50%
YP_001143866.1	*Aeromonas salmonicida*	50%
YP_854741.1	*Aeromonas hydrophila*	50%
ZP_01113275.1	*Reinekea sp. MED297*	50%
YP_004394511.1	*Aeromonas veronii*	50%
YP_004482887.1	*Marinomonas posidonica*	50%
ZP_06753531.1	*Simonsiella muelleri*	50%
YP_004314448.1	*Marinomonas mediterranea*	50%
YP_003913334.1	*Ferrimonas balearica*	50%
YP_001050781.1	*Shewanella baltica*	50%
YP_001093666.1	*Shewanella loihica*	50%
YP_002796370.1	*Laribacter hongkongensis*	50%
YP_525528.1	*Saccharophagus degradans*	50%
NP_967848.1	*Bdellovibrio bacteriovorus*	50%
YP_340711.1	*Pseudoalteromonas haloplanktis*	50%
YP_003931663.1	*Pantoea vagans*	49%
ZP_09757967.1	*Alishewanella jeotgali*	49%

Accession	Organism	Max ident
YP_003520842.1	*Pantoea ananatis*	49%
YP_001907224.1	*Erwinia tasmaniensis*	49%
YP_003531652.1	*Erwinia amylovora*	49%
ZP_01166607.1	*Oceanospirillum sp. MED92*	49%
ZP_09828983.1	*Pantoea stewartii*	49%
ZP_09987515.1	*Rheinheimera nanhaiensis*	49%
YP_002311434.1	*Shewanella piezotolerans*	49%
YP_002648366.1	*Erwinia pyrifoliae*	49%
YP_001758769.1	*Shewanella woody*	49%
YP_006009758.1	*Shewanella putrefaciens*	49%
ZP_09012845.1	*Commensalibacter intestini*	49%
YP_004434029.1	*Glaciecola sp. 4H-3-7+YE-5*	49%
YP_661403.1	*Pseudoalteromonas atlantica*	49%
YP_002894053.1	*Tolumonas auensis*	49%
YP_003626609.1	*Moraxella catarrhalis*	49%
YP_001475959.1	*Shewanella sediminis*	49%
NP_720477.1	*Shewanella oneidensis*	49%
ZP_01218901.1	*Photobacterium profundum*	49%
ZP_05886782.1	*Vibrio coralliilyticus*	49%
ZP_01085506.1	*Synechococcus sp. WH 5701*	49%
YP_944812.1	*Psychromonas ingrahamii*	49%
ZP_08068476.1	*Actinobacillus ureae*	49%
ZP_01308596.1	*Oceanobacter sp. RED65*	49%
ZP_07543350.1	*Actinobacillus pleuropneumoniae serovar*	49%
ZP_09864898.1	*Methylomicrobium album*	49%

```
>Kangiella_koreensis
MSNEVIKCKAAVAWEAGKPLSIEEVEVQPPQKGEVRVKIVATGVCHTDAFTLSGDDPEGVFPSILGHEGG
GIVESVGEGVTSVKPGDHVIPLYTPECGDCKFCLSGKTNLCQKIRETQGKGLMPDGTTRFSINGKPIYHY
MGTSTFSEYTVLPEISLAKVNPKAPLEEVCLLGCGVTTGMGAVMNTAKVEEGATVAIFGLGGIGLSAVIG
AVMAKASRIIAIDINESKFELAKKLGATDCVNPKDYDKPIQEVIVEMTDGGVDYSFECIGNVNVMRSALE
CCHKGWGESVIIGVAGAGQEISTRPFQLVTGRVWKGTAFGGVKGRSELPDYVERYLAGEFKLDDFITHTM
PLEKINDAFDLMHEGKSIRSVIHY

>Chromobacterium_violaceum
MSHDIIRCQAAVAWAAGQPLSIEEIEVHPPKAGEVRVKMVATGVCHTDAFTLSGADPEGVFPCILGHEGG
GVVESVGPGVTSVAVGDHVIPLYTPECRECKFCLSGKTNLCQKIRATQGKGLMPDGTSRFSKDGKPIYHY
MGTSTFSEYTVLPEISLAKVNKAAPLEEVCLLGCGVTTGMGAVVNTAKVKAGDNVAVFGLGGIGLSAIIG
ARMAGAGRIIGIDINEGKFELAKKLGATDCVNPNGFDKPIQDVIVEMTDGGVDFSFECIGNVKVMRAALE
CCHKGWGESVIIGVAGAGQEISTRPFQLVTGRVWRGSAFGGVRGRTELPEYVERYLKGEFRLDDFITHTM
PLERVNEAFDLMHEGKSIRSVIHYAAEPA

>Aeromonas_caviae
MAQVQSIKCKAAIAWGPGQPLSIEEVEVMPPQAGEVRVRIVATGVCHTDAFTLSGEDPEGVFPCILGHEG
GGIVESVGEGVTSVKVGDHVIPLYTPECGECKFCKSGKTNLCQKIRATQGKGLMPDGTTRFSKDGQPIYH
YMGTSTFSEYTVLPEISIAKVDPAAPLEEVCLLGCGVTTGIGAVMNTAKVKEGESVAIFGLGGIGLSAII
GARLAKAGRIIAIDINESKFELARKLGATDCINPNTFDKPIQEVIVEMTDGGVDFSFECIGNVKVMRAAL
ECCHKGWGESVIIGVAGAGQEISTRPFQLVTGRVWRGSAFGGVRGRSELPSYVQRYMQGEFKLDDFITHT
MPLEQINEAFDLMHEGKSIRTVIHY

>Marinobacter_manganoxydans
MTQTIKSKAAIAWGPKQPLSIEEVDVMPPQAGEVRIRVIASGVCHTDAFTLSGEDPEGIFPTILGHEGGG
IVESVGEGVTSLKVGDHVIPLYTPECGECKFCTSGKTNLCGKIRETQGKGLMPDGTTRFSLNGEPIYHYM
GCSTFSEYTVLPEISLAKVNKEAPLEEVCLLGCGVTTGMGAVMNTAKVEEGATVAIFGMGGIGLSAVIGA
TMAKASRIIAIDINESKFELARQLGATDCINPKDYDKPIQEVIVELTDGGVDYSFECIGNVDVMRSALEC
CHKGWGESVVIGVAGAGQEISTRPFQLVTGRVWKGSAFGGVKGRSELPGIVERYMQGEFKLNDFITHTMG
LEDINKAFDLMHEGKSIRTVIHYDK

>Alteromonas_sp. S89
MSAEPITCKAAVAWKAGEPLSIEEVVVAPPKAGEVRIRLLATGVCHTDAFTLSGADPEGVFPAILGHEGG
GVVESVGEGVTSVAVGDHVIPLYTPECGECKFCLSGKTNLCQKIRATQGKGLMPDGTSRFTVNGKPVFHY
MGTSTFSEYTVLPEISVAKVNKNAPLEEICLLGCGVTTGMGAVANTAKVEEGASVAVFGLGGIGLATIIG
ARLAKAGRIIAIDINEGKFELAKKLGATDCINPKSFDKPIQDVIVELTDGGVDYSFECIGNVDVMRSALE
CCHKGWGESVIIGVAGAGKEICTRPFQLVTGRVWRGTAFGGVKGRSQLPDYVERYLAGEFKLDDFITHTM
PLEKINEAFDLMHEGKSIRSVIHYA

>Marinobacter_aquaeolei
MAEIIKSKAAIAWGPGQPLSVEEVDVMPPKAGEVRIKVIATGVCHTDAFTLSGEDPEGNFPAILGHEGGG
IVEAIGEGVTSVAVGDHVIPLYTPECGECKFCLSGKTNLCGKIRETQGKGLMPDGTSRFYKDGQPIYHYM
GCSTFSEYTVLPEISLAKVNKEAPLEEVCLLGCGVTTGMGAVMNTAKVEEGATVAIFGLGGIGLSAIIGA
TMAKASRIIGIDINDSKFDLARQLGATDCINPKDYDKPIQEVIVELTDGGVDYSFECIGNVDVMRSALEC
CHKGWGESVIIGVAGAGQEISTRPFQLVTGRVWRGSAFGGVKGRSELPGIVERYLQGEFKLNDFITHTMG
LDDINEAFELMHEGKSIRSVIHFDK

>Marinobacter_hydrocarbonoclasticus
MAEMIKSKAAIAWGPGQPLSVEEVDVMPPKAGEVRIKVIATGVCHTDAFTLSGEDPEGNFPAILGHEGGG
IVEAIGEGVTSVAVGDHVIPLYTPECGECKFCLSGKTNLCGKIRETQGKGLMPDGTSRFYKDGQPIYHYM
GCSTFSEYTVLPEISLAKVNKEAPLEEVCLLGCGVTTGMGAVMNTAKVEEGATVAIFGLGGIGLSAIIGA
TMAKASRIIVIDINNSKFDLARQLGATDCINPNDYDKPIQEVIVELTDGGVDYSFECIGNVDVMRSALEC
CHKGWGESVIIGVAGAGQEISTRPFQLVTGRVWRGSAFGGVKGRSELPGIVERYLQGEFKLNDFITHTMG
LDDINEAFELMHEGKSIRSVIHFDK

>Shewanella_halifaxensis
MMTAKTIKSKAAVAWAVGEPLTMEIVDVMPPQKGEVRIKMIATGVCHTDAFTLSGDDPEGIFPCILGHEG
GGIVESIGEGVTSVKVGDHVIPLYTPECGECKYCKSGKTNLCQKIRETQGKGLMPDGTTRFSKDGVEIFH
YMGTSTFSEYTVLPEISLAKVNPDAPLEEVCLLGCGVTTGMGAVMNTAKVEEGSTVAVFGMGGIGLSAII
GAVMAKASRIIAIDINESKFELARKLGATDCINPKDYDKPIQEVIIELTDGGVDYSFECIGNIHVMRSAL
ECCHKGWGESVIIGVAGAGQEISTRPFQLVTGRVWKGSAFGGVKGRSELPEYVERYMAGEFKLDDFITHT
MGLEQVNDAFDLMHEGKSIRTVLHFGK
```

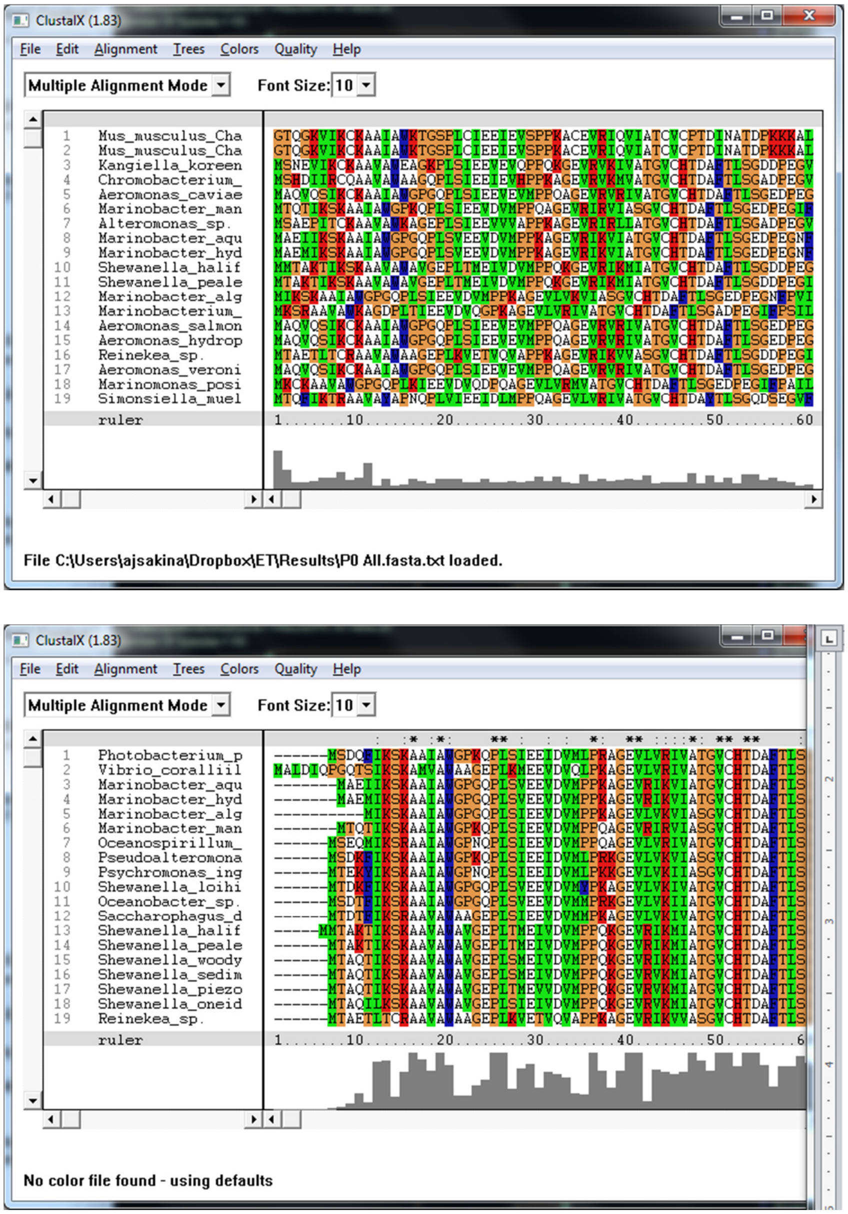
Figure 4.3: Sample FASTA of Some of the 48 Selected Similar Bacterial Sequences.

4.1.5 Multiple Sequence Alignment

Complete alignments of all 50 sequences from the 48 bacteria sequences and 2 chains of 1E3I are done using the ClustalX program. The alignment produces a .aln file that can be saved for future use. The output of the multiple sequences alignment is as Figure 4.4.

4.1.6 Construction of Phylogenetic Tree

A Phylip Tree .ph file is generated from the ClustalX post alignment. The .ph file is used to construct a phylogenetic tree in the PhyloDraw program. The Rectangular Cladogram of the constructed tree is as Figure 4.5. The tree is in a horizontal orientation and made up of two main ancestral nodes centred over their descendants.



Multiple Alignment Mode

Font Size: 10

```
1   Mus_musculus_Cha    GTQGKVIKCKAAIAWKTGSPLCIEEIEVSPPKACEVRIQVIATCVCPTDINATDPKKKAL
2   Mus_musculus_Cha    GTQGKVIKCKAAIAWKTGSPLCIEEIEVSPPKACEVRIQVIATCVCPTDINATDPKKKAL
3   Kangiella_koreen    MSNEVIKCKAAVAWEAGKPLSIEEVEVQPPQKGEVRVKIVATGVCHTDAFTLSGDDPEGV
4   Chromobacterium_    MSHDIIRCQAAVAWAAGQPLSIEEIEVHPPKAGEVRVKMVATGVCHTDAFTLSGADPEGV
5   Aeromonas_caviae    MAQVQSIKCKAAIAWGPGQPLSIEEVEVMPPQAGEVRVRIVATGVCHTDAFTLSGEDPEG
6   Marinobacter_man    MTQTIKSKAAIAWGPKQPLSIEEVDVMPPQAGEVRIRVIASGVCHTDAFTLSGEDPEGIF
7   Alteromonas_sp.     MSAEPITCKAAVAWKAGEPLSIEEVVVAPPKAGEVRIRLLATGVCHTDAFTLSGADPEGV
8   Marinobacter_aqu    MAEIIKSKAAIAWGPGQPLSVEEVDVMPPKAGEVRIKVIATGVCHTDAFTLSGEDPEGNF
9   Marinobacter_hyd    MAEMIKSKAAIAWGPGQPLSVEEVDVMPPKAGEVRIKVIATGVCHTDAFTLSGEDPEGNF
10  Shewanella_halif    MMTAKTIKSKAAVAWAVGEPLTMEIVDVMPPQKGEVRIKMIATGVCHTDAFTLSGDDPEG
11  Shewanella_peale    MTAKTIKSKAAVAWAVGEPLTMEIVDVMPPQKGEVRIKMIATGVCHTDAFTLSGDDPEGI
12  Marinobacter_alg    MIKSKAAIAWGPGQPLSIEEVDVMPPKAGEVLVKVIASGVCHTDAFTLSGEDPEGNFPVI
13  Marinobacterium_    MKSRAAVAWKAGDPLTIEEVDVQGPKAGEVLVRIVATGVCHTDAFTLSGADPEGIFPSIL
14  Aeromonas_salmon    MAQVQSIKCKAAIAWGPGQPLSIEEVEVMPPQAGEVRVRIVATGVCHTDAFTLSGEDPEG
15  Aeromonas_hydrop    MAQVQSIKCKAAIAWGPGQPLSIEEVEVMPPQAGEVRVRIVATGVCHTDAFTLSGEDPEG
16  Reinekea_sp.        MTAETLTCRAAVAWAAGEPLKVETVQVAPPKAGEVRIKVVASGVCHTDAFTLSGDDPEGI
17  Aeromonas_veroni    MAQVQSIKCKAAIAWGPGQPLSIEEVEVMPPQAGEVRVRIVATGVCHTDAFTLSGEDPEG
18  Marinomonas_posi    MKCKAAVAWGPGQPLKIEEVDVQDPQAGEVLVRMVATGVCHTDAFTLSGEDPEGIFPAIL
19  Simonsiella_muel    MTQFIKTRAAVAYAPNQPLVIEEIDLMPPQAGEVLVRIVATGVCHTDAYTLSGQDSEGVF
    ruler               1.......10........20........30........40........50........60
```

File C:\Users\ajsakina\Dropbox\ET\Results\P0 All.fasta.txt loaded.

Multiple Alignment Mode

Font Size: 10

```
                                   :  :* :*:     ** :   :  :    *:   ** ::::*:  ** **  :
1   Photobacterium_p    ------MSDQFIKSKAAIAWGPKQPLSIEEIDVMLPRAGEVLVRIVATGVCHTDAFTLS
2   Vibrio_coralliil    MALDIQPGQTSIKSKAMVAWAAGEPLKMEEVDVQLPKAGEVLVRIVATGVCHTDAFTLS
3   Marinobacter_aqu    -------MAEIIKSKAAIAWGPGQPLSVEEVDVMPPKAGEVRIKVIATGVCHTDAFTLS
4   Marinobacter_hyd    -------MAEMIKSKAAIAWGPGQPLSVEEVDVMPPKAGEVRIKVIATGVCHTDAFTLS
5   Marinobacter_alg    ----------MIKSKAAIAWGPGQPLSIEEVDVMPPKAGEVLVKVIASGVCHTDAFTLS
6   Marinobacter_man    -------MTQTIKSKAAIAWGPKQPLSIEEVDVMPPQAGEVRIRVIASGVCHTDAFTLS
7   Oceanospirillum_    ------MSEQMIKSRAAIAWGPNQPLSIEEVDVMPPQAGEVLVRIVATGVCHTDAFTLS
8   Pseudoalteromona    ------MSDKFIKSKAAIAWGPKQPLSIEEIDVMLPRKGEVLVKVIASGVCHTDAFTLS
9   Psychromonas_ing    ------MTEKYIKSKAAIAWGPNQPLSIEEIDVMLPKKGEVLVKIIASGVCHTDAFTLS
10  Shewanella_loihi    ------MTDKFIKSKAAIAWGPGQPLSVEEVDVMYPKAGEVLVKIVATGVCHTDAFTLS
11  Oceanobacter_sp.    ------MSDTFIKSKAAIAWGPGQPLSVEEVDVMMPRKGEVLVKIIATGVCHTDAFTLS
12  Saccharophagus_d    ------MTDTFIKSRAAVAWAAGEPLSIEEVDVMMPKAGEVLVKIVATGVCHTDAFTLS
13  Shewanella_halif    -----MMTAKTIKSKAAVAWAVGEPLTMEIVDVMPPQKGEVRIKMIATGVCHTDAFTLS
14  Shewanella_peale    ------MTAKTIKSKAAVAWAVGEPLTMEIVDVMPPQKGEVRIKMIATGVCHTDAFTLS
15  Shewanella_woody    ------MTAQTIKSKAAVAWAVGEPLSMEIVDVMPPQKGEVRIKMIATGVCHTDAFTLS
16  Shewanella_sedim    ------MTAQTIKSKAAVAWAVGEPLSMEIVDVMPPQKGEVRVKMIATGVCHTDAFTLS
17  Shewanella_piezo    ------MTAQTIKSKAAVAWAVGEPLTMEVVDVMPPQKGEVRVKMIATGVCHTDAFTLS
18  Shewanella_oneid    ------MTAQILKSKAAVAWAVGEPLSIEIVDVMPPQKGEVRVKMIATGVCHTDAFTLS
19  Reinekea_sp.        ------MTAETLTCRAAVAWAAGEPLKVETVQVAPPKAGEVRIKVVASGVCHTDAFTLS
    ruler               1.......10........20........30........40........50.......6
```

No color file found - using defaults

Figure 4.4: Pre Alignment (top) and Post Multiple Sequences Alignment (bottom) using ClustalX.

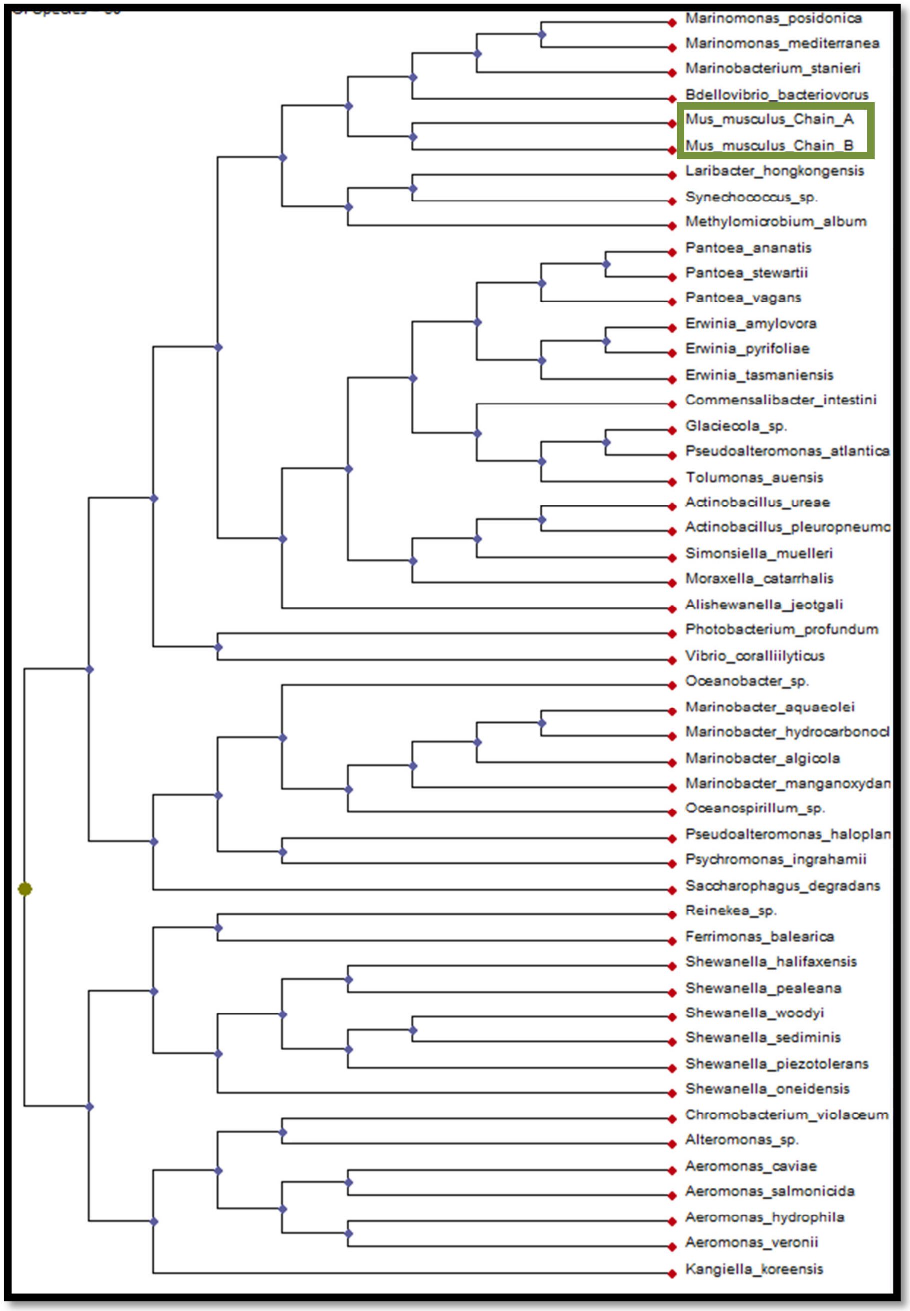


Figure 4.5: Rectangular Cladogram using PhyloDraw. The Mus musculus Query Sequence is Highlighted in Green.

4.2 TASK 2: IDENTIFICATION OF CONSENSUS SEQUENCES AND EVOLUTIONARY TRACE

4.2.1 Partitioning the Cladogram and Species Grouping

The partitions are made by making 5 vertical lines across the dendogram which are identified as P0, P1, P2, P3 and P4 as seen in Figure 4.6. The partition generates groups of an entire family that branches off from a node. The lowest partition i.e. P0 has smaller number of groups with bigger number of organisms in each group. The highest partition i.e. P4 has bigger number of groups but with smaller number of organisms in each group.

This analysis focuses on the groups with the star symbols beside their partition and group numbers as shown in Figure 4.7. The organisms which are being left as a single leaf (a species on their own with no other members) on each of the partition will be ignored and will not be used for further analysis. Only groups consisting of a family of organisms will be considered for analysis.

P0 branches out into two groups where Group 1 consists of 35 members and Group 2 has 15 members.

P1 is divided into 4 groups with 26 members in Group 1, 9 organisms in Group 2, 8 organisms in Group 3 and 7 organisms in Group 4.

P2 is made up of 6 groups with 24 members in Group 1, 2 organisms each in Group 2 and Group 4, 8 organisms in Group 3, 6 species all from the genus *Shewanella* in Group 5 and Group 6 consists of 6 organisms for of which are from the genus *Aeromonas*. Two organisms are being eliminated in P2 which are *Saccharophagus degradens* and *Kangiella koreensis* as they have branched out from the node individually as a single leaf on the tree.

P3 is originally made up of 7 groups. However, only the first two groups are being taken into consideration for further analysis. The rest of the organisms are being eliminated. Group 1 consists of 9 members and Group 2 15 organisms.

The final partition i.e. P4 has altogether 8 groups. The 3 groups with the most number of similar amino acid residues to the query sequence *Mus musculus* are being used for further analysis. Group 1 consists of 6 members with 5 species as there are two sequences for the *Mus musculus* species. Group 2 has 3 members and Group 3 with 14 species has the most members. The remaining organisms in the other remaining groups with lesser number of similar amino acid residues are being eliminated from the next step of analysis.

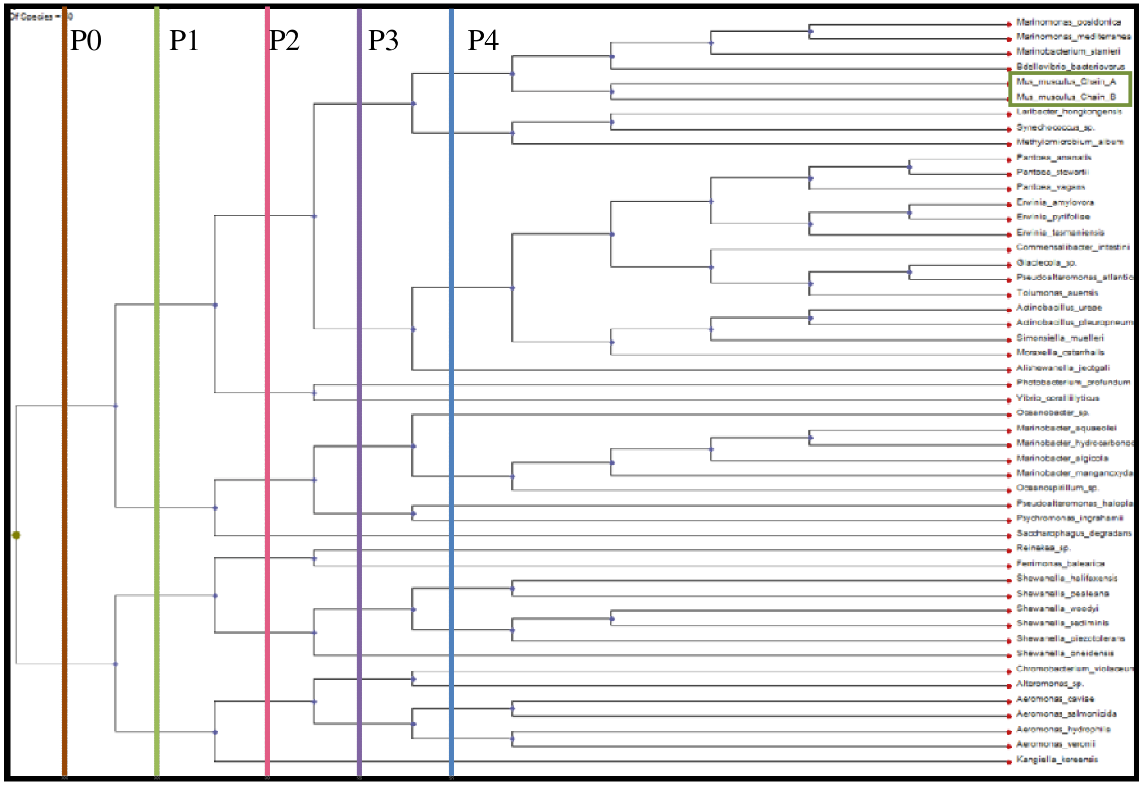


Figure 4.6: Rectangular Cladogram Partitioning. The Mus musculus Query Sequence is Highlighted in Green.

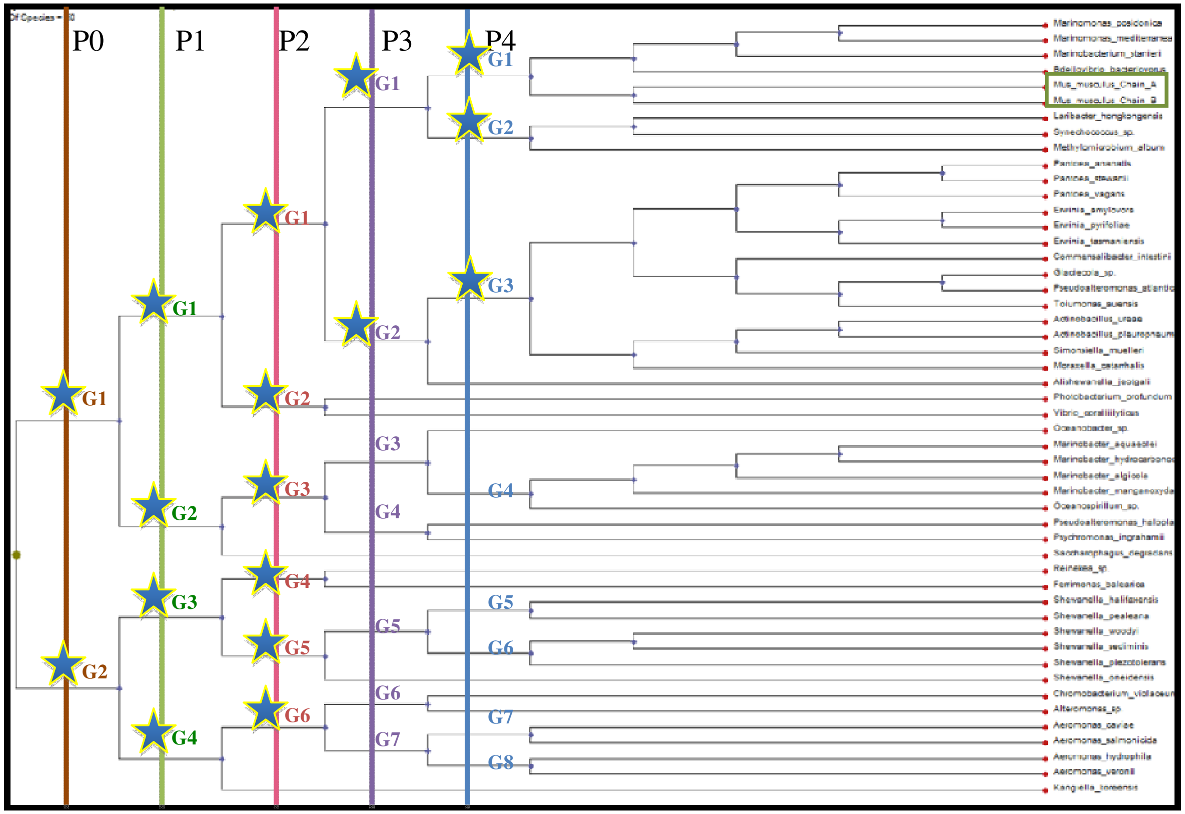


Figure 4.7: Rectangular Cladogram Grouping from the Partitions. The Stars show the Groupings from the Partitions.

4.2.2 Single Group Multiple Sequences Alignment

The members in each of the group are being aligned by partition to determine the conserved and neutral residues in the group. Figure 4.8 shows an example of the Partition P1 multiple sequences alignment results by group.

4.2.3 Identification of Consensus Sequences and Evolutionary Trace

For each partition, the results from the group alignment will be compared and a consensus is built. The consensus sequences consist of conserved residues, class-specific residues (if any) and neutral residues. The conserved residues are residues that exist in each of the organisms in the group being aligned. The class-specific residues are amino acid residues that are conserved within the group but differ between groups. The neutral residues are residues that are neither conserved within the group nor between the groups.

The conserved residues are being translated on the trace record as they are, the class-specific residues are translated as X and highlighted in green while the neutral residues are shown as dashes (-) in Figure 4.9.

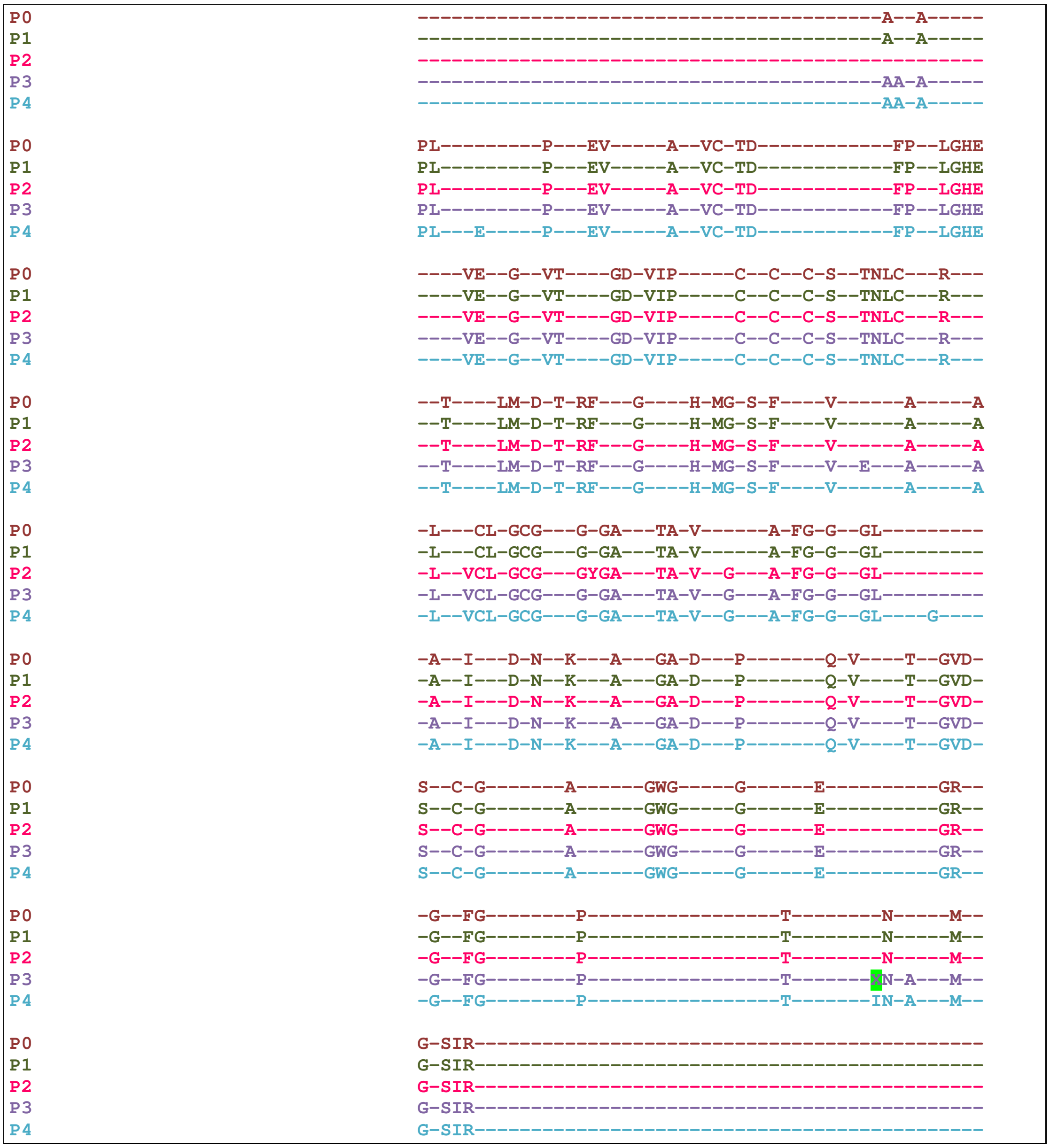
```
Marinomonas_posidonica              KGSAFGGVKGRSQLPGYVEDYMNGKIEIDPFVTHTMPLEQINEAFDLMHE 359
Marinomonas_mediterranea            KGSAFGGVKGRSQLPGYVEDYMNGKIEIDPFVTHTMSLEKINDAFDLMHE 359
Marinobacterium_stanieri            KGSAFGGVKGRSQLPGYVEDYMNGKIEIDPFITHTMGLEDINKAFDLMHE 359
Bdellovibrio_bacteriovorus          KGSAFGGVKGRTELPGYVEQYMSGEINIDDMVTFTMPLEDINKAFDYMHE 359
Mus_musculus_Chain_A                NGTFFGGWKSVDSVPNLVSDYKNKKFDLDLLVTHALPFESINDAIDLMKE 366
Mus_musculus_Chain_B                NGTFFGGWKSVDSVPNLVSDYKNKKFDLDLLVTHALPFESINDAIDLMKE 366
Laribacter_hongkongensis            RGSAFGGVRGRTELPAYVEKAQKGEIPLDTFITHTLPLEEINQAFELMHE 361
Synechococcus_sp.                   RGSAFGGVRGRTELPGYVERFQSGEIPLDTFITHTMPLEEINRAFELMHA 360
Methylomicrobium_album              RGSAFGGVHGRSELPGYVERAQRGEIPLDVFITHTLGLEDINQAFDLMHE 360
Pantoea_ananatis                    RGSAFGGVKGRSQLPGIVQDYLDGKFALNDFITHTMPLEEINDAFDLMHE 362
Pantoea_stewartii                   RGSAFGGVKGRSQLPGIVQDYLDGKFALNDFITHTMPLAEINDAFDLMHE 362
Pantoea_vagans                      RGSAFGGVKGRSQLPGIVQDYLDGKFALNDFITHTMPLEEINDAFDLMHE 362
Erwinia_amylovora                   RGSAFGGVKGRTQLPGIVERYMNGEFQLNDFITHNLPLEEINDAFELMHE 362
Erwinia_pyrifoliae                  RGSAFGGVKGRTQLPGIVERYMNGEFRLNDFITHNLPLEEINDAFELMHE 362
Erwinia_tasmaniensis                RGSAFGGVKGRTQLPGLVERYMGGEFQLNDFITHNLPLEQINDAFDLMHE 362
Commensalibacter_intestini          RGSAFGGVKGRSQLPNIVNDYLQGKFQLDDFITHEMPLDQINKAFDLMHD 362
Glaciecola_sp.                      RGSAFGGVKGRSQLPDYVQRYMDGEFELDTFITHTMQLEDINTAFDLMHE 362
Pseudoalteromonas_atlantica         RGSAFGGVKGRSQLPDYVQRYMDGEFELDTFITHTMPLEDINTAFDLMHE 362
Tolumonas_auensis                   KGSAFGGVKGRSQLPGIVEQYMNGEFELDTFITHTMGLDDINHAFDLMHE 362
Actinobacillus_ureae                RGSAFGGVKGRTELPGIIDQFMKGEFKLRDFITHTMPLEDINKAFDLMHQ 362
Actinobacillus_pleuropneumonia      RGSAFGGVKGRTELPGIIDQFMKGEFKLRDFITHTMPLEDINKAFDLMHE 362
Simonsiella_muelleri                RGSAFGGYKGRSELPDLIDQYQHGEFKLSDFITHTMPLEDINNAFDLMHE 363
Moraxella_catarrhalis               RGSAFGDVKGRSELPGIVSQYMQGDFALSDFITHTMPLDQINAAFDLMHE 362
Alishewanella_jeotgali              RGTAFGGVKGRSELPGYVDRYLNGEFELDTFITHTMPLEDINKAFDLMHE 362
Photobacterium_profundum            RGSAFGGVKGRSELPEIVERYMAGEFALDDFITHTMGLEGINDAFDLMHE 364
Vibrio_coralliilyticus              RGSAFGGVKGRSELPEIVERYMAGEFGLQEFITHTMGLDAVNDAFDLMHE 370
P1 Group 1                          -G--FG--------P-----------------T--------N-A---M--
                                    .*: **. :.  .:*  :.     .: :  ::*. : :  :* *:: *:

Oceanobacter_sp.                    RGSAFGGVKGRSELPGIVEKYLAGEFKLNDFITHTMGLEDINEAFDLMHE 364
Marinobacter_aquaeolei              RGSAFGGVKGRSELPGIVERYLQGEFKLNDFITHTMGLDDINEAFELMHE 363
Marinobacter_hydrocarbonoclast      RGSAFGGVKGRSELPGIVERYLQGEFKLNDFITHTMGLDDINEAFELMHE 363
Marinobacter_algicola               KGSAFGGVKGRSELPGIVERYLQGEFKLNDFITHTMGLDDINEAFELMHE 360
Marinobacter_manganoxydans          KGSAFGGVKGRSELPGIVERYMQGEFKLNDFITHTMGLEDINKAFDLMHE 363
Oceanospirillum_sp.                 RGTAFGGVKGRSELPEIVERYMAGEFKLNDFITHTMGLDKINEAFDLMHE 364
Pseudoalteromonas_haloplanktis      RGSAFGGVKGRSELPDYVERYLAGEFKLSDFITHTMGLEDINESFDLMRR 364
Psychromonas_ingrahamii             KGSAFGGVKGRTELPDYVERYLQGEFKLSDFITHTMPLEDVNEAFELMHK 364
Saccharophagus_degradans            RGTAFGGVKGRSELPGIVEQYLAGDFKLDDFITHTMGLEDINTAFDLMHH 364
P1 Group 2                          -G-AFGGVKGR-ELP--VE-Y--G-FKL-DFITHTM-L---N--F-LM--
                                    :*:********:***  **:*: *:***.******* *:.:* :*:**:.

Reinekea_sp.                        RGTAFGGVKGRSELPSYVERYLDGEFKLSDFITHTMPLDEINEAFDLMHE 364
Ferrimonas_balearica                RGSAFGGVKGRSELPQYVERYLAGEFKLDDFITHTMGLDKINDAFDLMHQ 365
Shewanella_halifaxensis             KGSAFGGVKGRSELPEYVERYMAGEFKLDDFITHTMGLEQVNDAFDLMHE 365
Shewanella_pealeana                 KGSAFGGVKGRSELPEYVERYMAGEFKLDDFITHTMGLEQVNDAFDLMHE 364
Shewanella_woodyi                   KGSAFGGVKGRSELPEYVERYMAGEFKLNDFITHTMGLEQVNEAFDLMHE 364
Shewanella_sediminis                KGSAFGGVKGRSELPEYVERYMAGEFKLNDFITHTMGLEQVNDAFDLMHE 364
Shewanella_piezotolerans            KGSAFGGVKGRSELPEYVERYMAGEFKLNDFITHTMGLDQVNEAFDLMHE 364
Shewanella_oneidensis               KGSAFGGVKGRSELPEYVERYLAGEFKLSDFITHTMSLEQVNDAFDLMHQ 364
P1 Group 3                          -G-AFGGVKGRSELP-YVERY--GEFKL-DFITHTM-L---N-AFDLMH-
                                    :*:************.*****: *****.******* *:::*:******:

Chromobacterium_violaceum           RGSAFGGVRGRTELPEYVERYLKGEFRLDDFITHTMPLERVNEAFDLMHE 364
Alteromonas_sp.                     RGTAFGGVKGRSQLPDYVERYLAGEFKLDDFITHTMPLEKINEAFDLMHE 364
Aeromonas_caviae                    RGSAFGGVRGRSELPSYVQRYMQGEFKLDDFITHTMPLEQINEAFDLMHE 365
Aeromonas_salmonicida               RGSAFGGVRGRSELPSYVQRYMQGEFKLDDFITHTMPLEQINEAFELMHE 365
Aeromonas_hydrophila                RGSAFGGVRGRSELPSYVQRYMQGEFKLDDFITHTMGLEQINEAFDLMHE 365
Aeromonas_veronii                   RGSAFGGVRGRSELPSYVQRYMQGEFRLDDFITHTMGLEQINEAFELMHQ 365
Kangiella_koreensis                 KGTAFGGVKGRSELPDYVERYLAGEFKLDDFITHTMPLEKINDAFDLMHE 364
P1 Group 4                          -G-AFGGV-GR--LP-YV-RY--GEF-LDDFITHTM-LE--N-AF-LMH-
                                    :*:*****:**::**.**:**: ***:********* **::*:**:***:
```

Figure 4.8: Sample Multiple Sequences Alignment of Partition P1 by Group.



```
P0          -----------------------------------------A--A-----
P1          -----------------------------------------A--A-----
P2          --------------------------------------------------
P3          -----------------------------------------AA-A-----
P4          -----------------------------------------AA-A-----

P0          PL---------P---EV-----A--VC-TD------------FP--LGHE
P1          PL---------P---EV-----A--VC-TD------------FP--LGHE
P2          PL---------P---EV-----A--VC-TD------------FP--LGHE
P3          PL---------P---EV-----A--VC-TD------------FP--LGHE
P4          PL---E-----P---EV-----A--VC-TD------------FP--LGHE

P0          ----VE--G--VT----GD-VIP-----C--C--C-S--TNLC---R---
P1          ----VE--G--VT----GD-VIP-----C--C--C-S--TNLC---R---
P2          ----VE--G--VT----GD-VIP-----C--C--C-S--TNLC---R---
P3          ----VE--G--VT----GD-VIP-----C--C--C-S--TNLC---R---
P4          ----VE--G--VT----GD-VIP-----C--C--C-S--TNLC---R---

P0          --T----LM-D-T-RF---G----H-MG-S-F----V------A-----A
P1          --T----LM-D-T-RF---G----H-MG-S-F----V------A-----A
P2          --T----LM-D-T-RF---G----H-MG-S-F----V------A-----A
P3          --T----LM-D-T-RF---G----H-MG-S-F----V--E---A-----A
P4          --T----LM-D-T-RF---G----H-MG-S-F----V------A-----A

P0          -L---CL-GCG---G-GA---TA-V------A-FG-G--GL---------
P1          -L---CL-GCG---G-GA---TA-V------A-FG-G--GL---------
P2          -L--VCL-GCG---GYGA---TA-V--G---A-FG-G--GL---------
P3          -L--VCL-GCG---G-GA---TA-V--G---A-FG-G--GL---------
P4          -L--VCL-GCG---G-GA---TA-V--G---A-FG-G--GL----G----

P0          -A--I---D-N--K---A---GA-D---P-------Q-V----T--GVD-
P1          -A--I---D-N--K---A---GA-D---P-------Q-V----T--GVD-
P2          -A--I---D-N--K---A---GA-D---P-------Q-V----T--GVD-
P3          -A--I---D-N--K---A---GA-D---P-------Q-V----T--GVD-
P4          -A--I---D-N--K---A---GA-D---P-------Q-V----T--GVD-

P0          S--C-G-------A------GWG-----G------E----------GR--
P1          S--C-G-------A------GWG-----G------E----------GR--
P2          S--C-G-------A------GWG-----G------E----------GR--
P3          S--C-G-------A------GWG-----G------E----------GR--
P4          S--C-G-------A------GWG-----G------E----------GR--

P0          -G--FG--------P-----------------T--------N-----M--
P1          -G--FG--------P-----------------T--------N-----M--
P2          -G--FG--------P-----------------T--------N-----M--
P3          -G--FG--------P-----------------T-------XN-A---M--
P4          -G--FG--------P-----------------T-------IN-A---M--

P0          G-SIR---------------------------------------------
P1          G-SIR---------------------------------------------
P2          G-SIR---------------------------------------------
P3          G-SIR---------------------------------------------
P4          G-SIR---------------------------------------------
```

Figure 4.9: Consensus Sequences for Each of the Partitions.

4.3 TASK 3: PROTEIN MAPPING

4.3.1 Secondary Structure of 1E3I

Figure 4.10 shows the Amino Acid residues in ball and stick surrounding the NAD and CXF ligands located within 5 Armstrong of the ligands. The two zinc atoms can be seen in green where one of which is in the middle of the protein located closer to the ligands.

4.3.2 The Zinc Atom, NAD Ligand and CXF Ligand with Surrounding Residues

Figure 4.11 shows the NAD and CXF ligands in CPK. One zinc atom can be seen in gold which is located closer to the ligands. The rest of the ball and stick molecules in green, red, blue, yellow and orange are amino acid residues located within 5 Armstrong of the ligands.

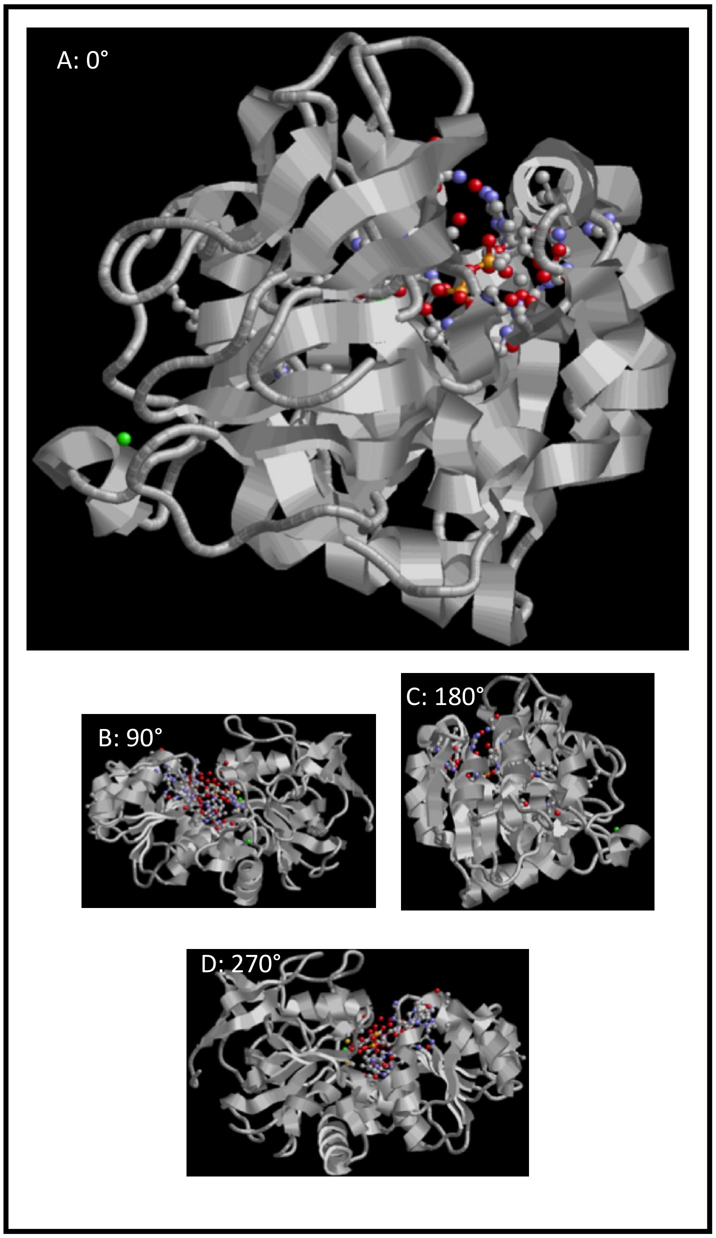


Figure 4.10: Secondary Structure of 1E3I with Zinc Atoms, NAD Ligand and CXF Ligand with Surrounding Residues on Consecutive 90° Rotations about the X-axis.

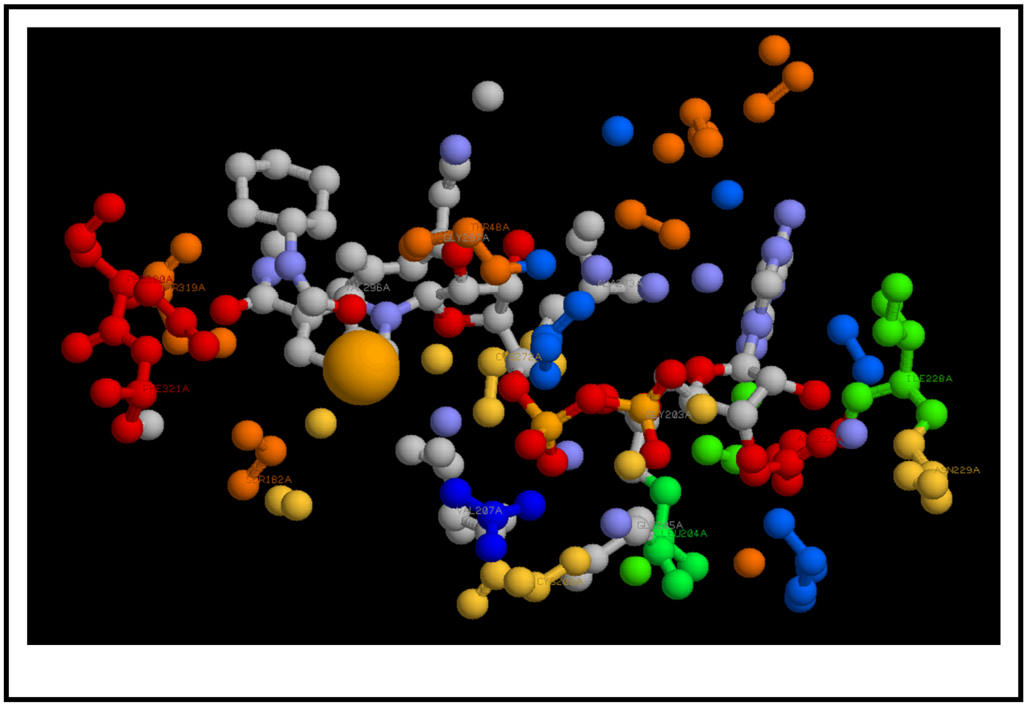


Figure 4.11: Zinc Atoms, NAD Ligand and CXF Ligand with Surrounding Residues.

4.3.3 Comparison of Trace Analysis, Rasmol and Active Sites Identified from PDB Sum

Table 4.2 and Table 4.3 show the comparison of the conserved residues identified from the partitions P0 and P4 respectively of the evolutionary trace analysis conserved residues, molecules from the Rasmol analysis and active site residues from PDB SUM.

The amino acid residues identified from the Rasmol analysis and exist in the evolutionary trace as conserved residues are highlighted in green. The amino acid residues which have been identified as active site residues in PDB SUM and exist in the evolutionary trace as conserved residues are highlighted in blue. If the amino acid residue exists in all three columns, the residue in the evolutionary trace column will be highlighted in yellow.

In partition P0, 109 residues out of the 376 residues or 29.8% residues are fully conserved throughout the whole group of 49 species. The conserved residues are mostly clustered in big groups, some are scattered in smaller groups and very few are scattered as single residues.

11 out of 23 active site residues identified using PDB SUM are fully conserved from the evolutionary trace analysis.

Table 4.2: Comparison of P0 Trace Analysis Conserved Residues, Rasmol Analysis and PDB SUM Active Site Residues. Green Shows Amino Acid Residues that Exist in Both Rasmol and are Evolutionary Trace Conserved residues, Blue shows the Amino Acid are PDB SUM Active Site Residues and Evolutionary Trace Conserved Residues, and Yellow Shows the Amino Acid Residues are Found in Rasmol, are PDB SUM Active Residues and Evolutionary Trace Conserved Residues.

Trace	Rasmol	PDB SUM
A11		
A14		
P20		
L21		
P31		
E35		
V36		
A42		
V45		
C46	CYS46	
	HIS47	HIS A 47
T48	THR48	THR A 48
D49		
F61		
P62		
L65		
G66		
H67		
E68		
V73		
E74		
G77		
V80		
T81		
G86		
D87		
V89		
I90		
P91		
C97		
C100		
C103		
S105		
T108		
N109		
L110		
C111		
R115		
T121		
L126		
M127		
D129		
T131		
R133		
F134		
G138		
H143		
M145		
G146		
S148		
F150		
V155		
A162		
A168		
L170		
C174		
L175		
G177		
C178		CYS A 178
G179		
	SER182	SER A 182
G183		
G185		
A186		
T190		
A191		
V193		
A200		
F202		
G203	GLY203	GLY A 203
	LEU204	LEU A 204
G205	GLY205	GLY A 205
	CYS206	CYS A 206
	VAL207	
G208		
L209		
A220		
I223		
D227	ASP227	ASP A 227
	ILE228	ILE A 228
N229	ASN229	ASN A 229
K232		LYS A 232
A236		
G240		
A241		
D243		
P247		
Q255		
V257		
T262		
G265		
V266		
D267		
S269		
C272	CYS272	CYS A 272
	ALA273	
G274		
		THR A 275
		GLN A 277
		THR A 278
A282		
G289		
W290		
G291		
	VAL296	
G297	GLY297	GLY A 297
	ALA298	ALA A 298
E304		
G313		
R314		
G318		
	THR319	THR A 319
	PHE320	PHE A 320
F321	PHE321	PHE A 321
G322		
P331		
T349		
N358		
M364		
G367		
S369		
I370		
R371		ARG A 371
		ASP B 362

Table 4.3: Comparison of P4 Trace Analysis Conserved Residues, Rasmol Analysis and PDB SUM Active Site Residues. Green Shows Amino Acid Residues that Exist in Both Rasmol and are Evolutionary Trace Conserved residues, Blue shows the Amino Acid are PDB SUM Active Site Residues and Evolutionary Trace Conserved Residues, and Yellow Shows the Amino Acid Residues are Found in Rasmol, are PDB SUM Active Residues and Evolutionary Trace Conserved Residues.

Trace	Rasmol	PDB SUM
A11		
A12		
A14		
P20		
L21		
E25		
P31		
E35		
V36		
A42		
V45		
C46	CYS46	
	HIS47	HIS A 47
T48	THR48	THR A 48
D49		
F61		
P62		
L65		
G66		
H67		
E68		
V73		
E74		
G77		
V80		
T81		
G86		
D87		
V89		
I90		
P91		
C97		
C100		
C103		
S105		
T108		
N109		
L110		
C111		
R115		
T121		
L126		
M127		
D129		
T131		
R133		

Trace	Rasmol	PDB SUM
F134		
G138		
H143		
M145		
G146		
S148		
F150		
V155		
A162		
A168		
L170		
V173		
C174		
L175		
G177		
C178		CYS A 178
G179		
	SER182	SER A 182
G183		
G185		
A186		
T190		
A191		
V193		
A200		
F202		
G203	GLY203	GLY A 203
	LEU204	LEU A 204
G205	GLY205	GLY A 205
	CYS206	CYS A 206
	VAL207	
G208		
L209		
G214		
A220		
I223		
D227	ASP227	ASP A 227
	ILE228	ILE A 228
N229	ASN229	ASN A 229
K232		LYS A 232
A236		
G240		
A241		
D243		
P247		
Q255		

Trace	Rasmol	PDB SUM
V257		
T262		
G265		
V266		
D267		
S269		
C272	CYS272	CYS A 272
	ALA273	
G274		
		THR A 275
		GLN A 277
		THR A 278
A282		
G289		
W290		
G291		
	VAL296	
G297	GLY297	GLY A 297
	ALA298	ALA A 298
E304		
G313		
R314		
G318		
	THR319	THR A 319
	PHE320	PHE A 320
F321	PHE321	PHE A 321
G322		
P331		
T349		
I357		
N358		
A360		
M364		
G367		
S369		
I370		
R371		ARG A 371
		ASP B 362

4.3.4 Mapping of Conserved Residues on the Protein Structure

The conserved residues appear to be dispersed in the protein structure. The residues are mixtures of big and small clusters spread throughout the surface of the protein. The conserved residues can be seen on Figure 4.12 in green.

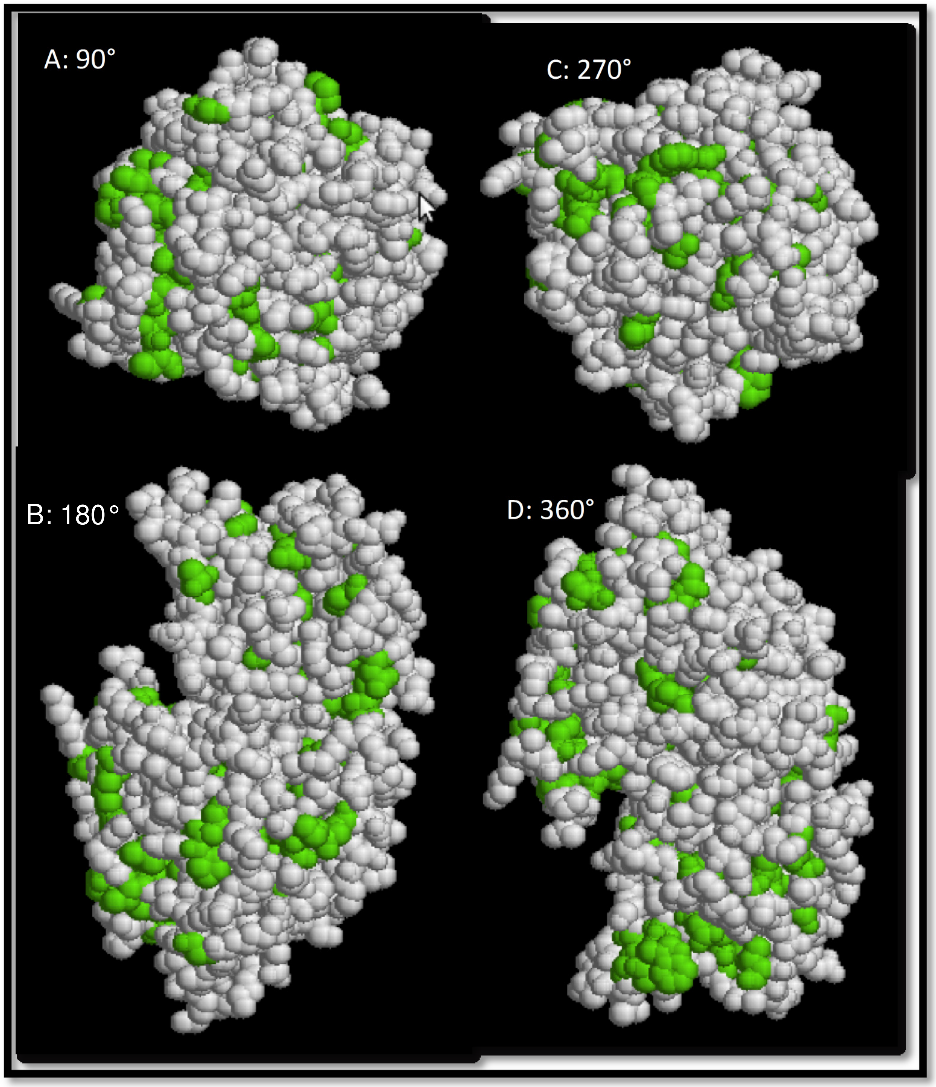


Figure 4.12: Conserved Residues Mapped on the Protein Surface on Consecutive 90° Rotations About the X-axis.

5 CHAPTER 5: DISCUSSION

5.1 TASK 1: DETERMINING THE PHYLOGENETIC RELATIONSHIP OF ALCOHOL DEHYDROGENASE

The RCSB PDB website held a huge data bank of proteins with plenty of information on the protein of the organism being studied. The 1E3I crystal structure of mouse class II alcohol dehydrogenase (ADH) revealed determinants of substrate specificity and catalytic efficiency of the protein. The study was done by Svensson, Hoeoeg, Schneider and Sandalova and published in the year 2000. Other related PDB entries were the 1CDO Alcohol Dehydrogenase Complexed with Nicotinamide Adenine Dinucleotide and Zinc, 1E3L P47H Mutant of Mouse Class II Alcohol Dehydrogenase Complex with NADH and 1E3E Mouse Class II Alcohol Dehydrogenase Complex with NADH.

The Clustal group of program has a general use of global multiple sequence alignment programs for both DNA and proteins. The alignment of the proteins produces biologically important multiple sequence alignments of divergent sequences. This enables the researcher to identify the similarities and differences from the group of sequences. The program calculates the best match for the selected sequences and lines them up in the window. The Clustal group of program also has a build in viewing feature to determine evolutionary relationships of the organisms being aligned by building Cladograms or Phylograms. The basic information the Clustal program provides is identification of conserved sequence regions. The results of the consensus being generated are symbols representing the degree of conservation observed in each column. The star symbol “*” means that the residues in that column are identical in all sequences in the alignment, the colon symbol “:” means that conserved substitutions have been observed and full stop symbol “.” means that semi-conserved substitutions are observed.

The phylogenetic tree was constructed using the PhyloDraw program that worked as a tool for generating phylogenetic trees. The PhyloDraw program supports a variety of multialignment programs (Clustal-W, pairwise distance matrix, Dialign2 and Phylip format) and visualizes many types of tree diagrams. The rectangular Cladogram being built showed relations among the organisms containing ADH being studied. The Cladogram uses lines that branch off in different directions ending at group of organisms. The lines traced back to where the organisms branch off and the branching off pointed to where a common ancestor was believed to have existed.

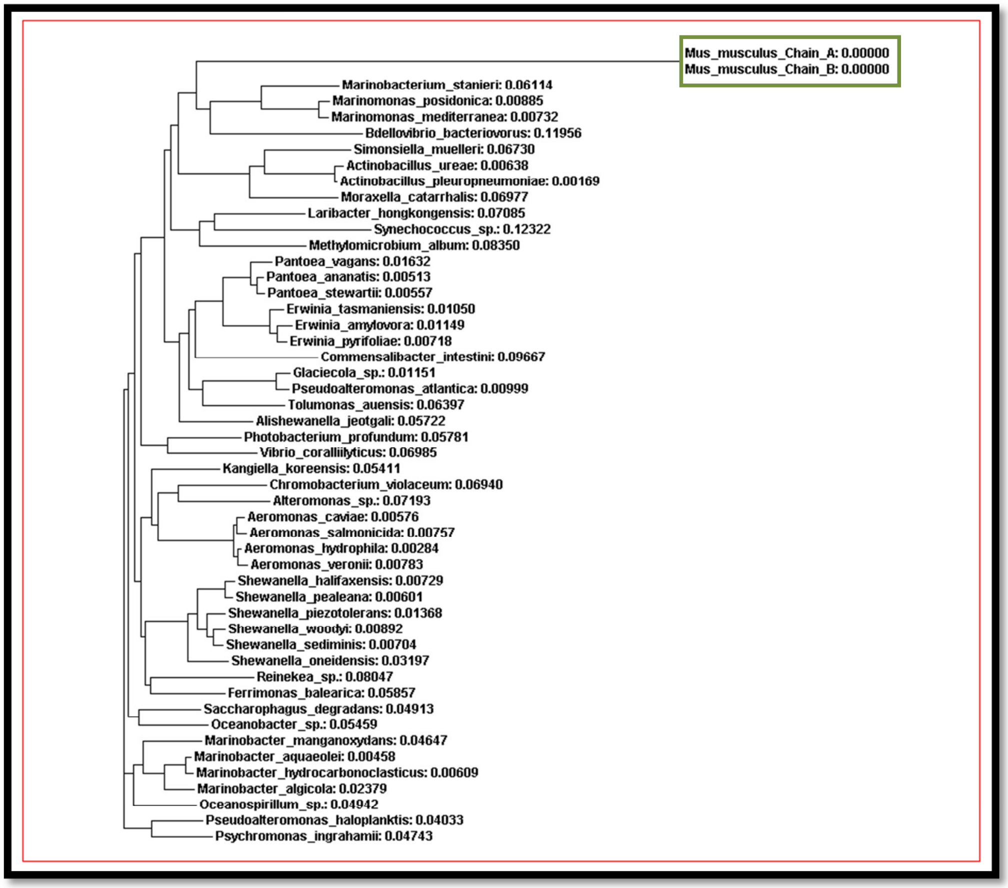

Mus_musculus_Chain_A: 0.00000
Mus_musculus_Chain_B: 0.00000
Marinobacterium_stanieri: 0.06114
Marinomonas_posidonica: 0.00885
Marinomonas_mediterranea: 0.00732
Bdellovibrio_bacteriovorus: 0.11956
Simonsiella_muelleri: 0.06730
Actinobacillus_ureae: 0.00638
Actinobacillus_pleuropneumoniae: 0.00169
Moraxella_catarrhalis: 0.06977
Laribacter_hongkongensis: 0.07085
Synechococcus_sp.: 0.12322
Methylomicrobium_album: 0.08350
Pantoea_vagans: 0.01632
Pantoea_ananatis: 0.00513
Pantoea_stewartii: 0.00557
Erwinia_tasmaniensis: 0.01050
Erwinia_amylovora: 0.01149
Erwinia_pyrifoliae: 0.00718
Commensalibacter_intestini: 0.09667
Glaciecola_sp.: 0.01151
Pseudoalteromonas_atlantica: 0.00999
Tolumonas_auensis: 0.06397
Alishewanella_jeotgali: 0.05722
Photobacterium_profundum: 0.05781
Vibrio_coralliilyticus: 0.06985
Kangiella_koreensis: 0.05411
Chromobacterium_violaceum: 0.06940
Alteromonas_sp.: 0.07193
Aeromonas_caviae: 0.00576
Aeromonas_salmonicida: 0.00757
Aeromonas_hydrophila: 0.00284
Aeromonas_veronii: 0.00783
Shewanella_halifaxensis: 0.00729
Shewanella_pealeana: 0.00601
Shewanella_piezotolerans: 0.01368
Shewanella_woodyi: 0.00892
Shewanella_sediminis: 0.00704
Shewanella_oneidensis: 0.03197
Reinekea_sp.: 0.08047
Ferrimonas_balearica: 0.05857
Saccharophagus_degradans: 0.04913
Oceanobacter_sp.: 0.05459
Marinobacter_manganoxydans: 0.04647
Marinobacter_aquaeolei: 0.00458
Marinobacter_hydrocarbonoclasticus: 0.00609
Marinobacter_algicola: 0.02379
Oceanospirillum_sp.: 0.04942
Pseudoalteromonas_haloplanktis: 0.04033
Psychromonas_ingrahamii: 0.04743

Figure 5.1: Phylogram as a Result of Multiple Sequence Alignment in ClustalW. The Mus musculus Query Sequence is Highlighted in Green.

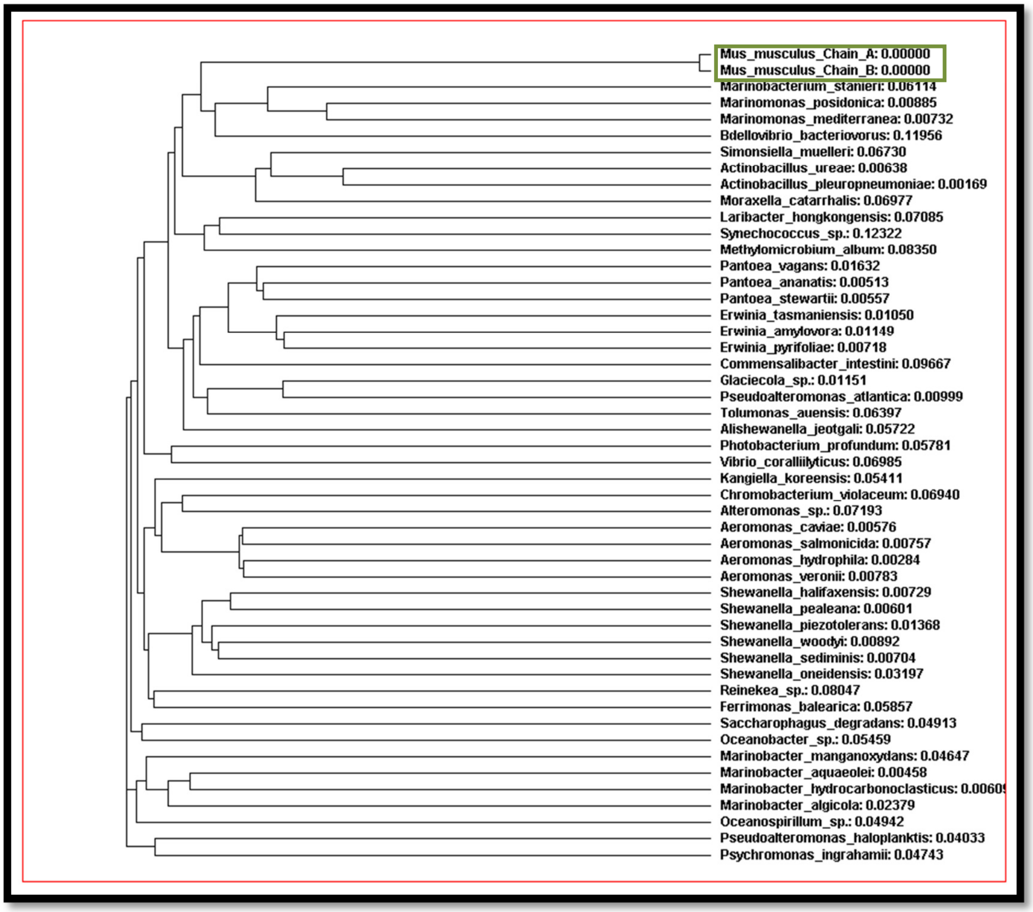

Mus_musculus_Chain_A: 0.00000
Mus_musculus_Chain_B: 0.00000
Marinobacterium_stanieri: 0.06114
Marinomonas_posidonica: 0.00885
Marinomonas_mediterranea: 0.00732
Bdellovibrio_bacteriovorus: 0.11956
Simonsiella_muelleri: 0.06730
Actinobacillus_ureae: 0.00638
Actinobacillus_pleuropneumoniae: 0.00169
Moraxella_catarrhalis: 0.06977
Laribacter_hongkongensis: 0.07085
Synechococcus_sp.: 0.12322
Methylomicrobium_album: 0.08350
Pantoea_vagans: 0.01632
Pantoea_ananatis: 0.00513
Pantoea_stewartii: 0.00557
Erwinia_tasmaniensis: 0.01050
Erwinia_amylovora: 0.01149
Erwinia_pyrifoliae: 0.00718
Commensalibacter_intestini: 0.09667
Glaciecola_sp.: 0.01151
Pseudoalteromonas_atlantica: 0.00999
Tolumonas_auensis: 0.06397
Alishewanella_jeotgali: 0.05722
Photobacterium_profundum: 0.05781
Vibrio_coralliilyticus: 0.06985
Kangiella_koreensis: 0.05411
Chromobacterium_violaceum: 0.06940
Alteromonas_sp.: 0.07193
Aeromonas_caviae: 0.00576
Aeromonas_salmonicida: 0.00757
Aeromonas_hydrophila: 0.00284
Aeromonas_veronii: 0.00783
Shewanella_halifaxensis: 0.00729
Shewanella_pealeana: 0.00601
Shewanella_piezotolerans: 0.01368
Shewanella_woodyi: 0.00892
Shewanella_sediminis: 0.00704
Shewanella_oneidensis: 0.03197
Reinekea_sp.: 0.08047
Ferrimonas_balearica: 0.05857
Saccharophagus_degradans: 0.04913
Oceanobacter_sp.: 0.05459
Marinobacter_manganoxydans: 0.04647
Marinobacter_aquaeolei: 0.00458
Marinobacter_hydrocarbonoclasticus: 0.0060
Marinobacter_algicola: 0.02379
Oceanospirillum_sp.: 0.04942
Pseudoalteromonas_haloplanktis: 0.04033
Psychromonas_ingrahamii: 0.04743

Figure 5.2: Cladogram as a Result of Multiple Sequence Alignment in ClustalW. The Mus musculus Query Sequence is Highlighted in Green.

5.2 TASK 2: IDENTIFICATION OF CONSENSUS SEQUENCES AND EVOLUTIONARY TRACE

The consensus sequences are calculated order of most frequent amino acid residues found at each position in a sequence alignment. The results of the multiple sequence alignment of related sequences are compared to each other. A known conserved sequence set is represented by a consensus sequence. Generally observed supersecondary protein structures or amino acid motifs are most of the time formed by conserved sequences. The comparison of interest in this task is for each group within the partition. The consensus information is important in the function of enzymes and coding genes. There are several coding region recognition programs available some of which are GenView, SpliceView and ORFGene. The consensus sequences in this task were done by submitting the sequences from each group using ClustalW and manually identifying the consensus sequences.

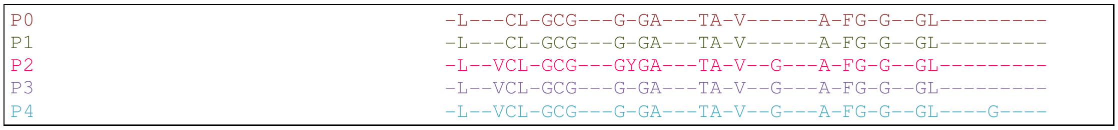
The evolutionary trace analysis was done for identification of conserved, class-specific and neutral residues between groups in a partition. The evolutionary trace would be able to predict the functional site by identifying a cluster of evolutionary important residues on the surface of the protein. 100 out of the 376 amino acid residues are fully conserved in partitions P0 and P1 and that made up of about 26.6% of the total residues. Partition P3 had about 27.7% conserved residues made up of 103 fully conserved residues and one class-specific residue. Partition P3 class-specific residue is at position 349 where the organisms in Group 1 had the Glutamic Acid and the amino acid was substituted by the Leucine in Group 2 organisms. The Leucine has a chemical formula of $HO_2CCH(NH_2)CH_2CH(CH_3)_2$. The Leucine is classified as a hydrophobic amino acid due to its aliphatic isobutyl side chain and it is an essential amino acid which is encoded by six codons (UUG, UUA, CUA, CUC, CUG and CUU). On the other hand, the Glutamic Acid is a non-essential amino acid with a chemical formula of $C_5H_9NO_4$. It is one of the 20-22 proteinogenic amino acids encoded by two codons (GAG and GAA). The carboxylate anions and salts of glutamic acid are known as glutamates. In neuroscience, glutamate is a vital neurotransmitter in long-term potentiation and is essential for memory and learning. The P4 partition had 105 fully conserved residues throughout the group which consisted of about 27.9% of the total amino acid residues.

All four amino acid residues ligated with zinc ions (Cys 46, Thr 48, His 67 and Cys 174) as reported in Ryan *et al* 1999 were identified as conserved residues in the evolutionary trace analysis done.

5.3 TASK 3: PROTEIN MAPPING

The protein mapping aims to map conserved residues to the surface of the protein. The analysis would help researchers identify residues critical for DNA binding where information on interactions between an enzyme and its substrates can be obtained.

The protein residues that are critical for structure and function are expected to be conserved throughout evolution. The conserved amino acid residues were seen to be clustered in groups of big and small residues. Only a few conserved amino acid residues were individually scattered.



```
P0                                  -L---CL-GCG---G-GA---TA-V------A-FG-G--GL---------
P1                                  -L---CL-GCG---G-GA---TA-V------A-FG-G--GL---------
P2                                  -L--VCL-GCG---GYGA---TA-V--G---A-FG-G--GL---------
P3                                  -L--VCL-GCG---G-GA---TA-V--G---A-FG-G--GL---------
P4                                  -L--VCL-GCG---G-GA---TA-V--G---A-FG-G--GL----G----
```

Figure 5.3: Sample Section of Conserved Amino Acid Residues Clusters.

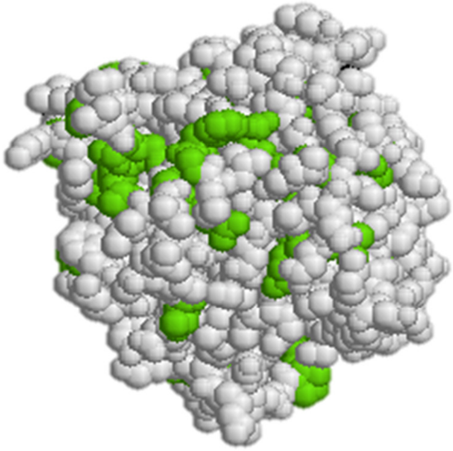


Figure 5.4: Amino Acid Conserved Residues Form Clusters on the Protein Surface.

6 CHAPTER 6: SUMMARY

The study was on the zinc metalloenzyme alcohol dehydrogenase evolutionary trace. An evolutionary trace analysis of the alcohol dehydrogenase enzyme was done using the software Rasmol, Clustal programs and PhyloDraw.

The FASTA sequence with the PDB ID 1E3I was used to study the evolutionary analysis by comparing with 48 other bacteria sequences. The 1E3I sequence had a length of 376 amino acid residues with two identical chains A and B. The BLAST search identified the bacteria sequences to be similar to the 1E3I sequence at the range of 48% to 52% similarity. The output from the multiple sequence alignment done helped to in construction of a phylogenetic tree. The rectangular Cladogram was used for analysis.

The Cladogram was partitioned into 5 vertical lines across the Cladogram. The partition generates groups of an entire family that branches off from a node. The earlier partition i.e. P0 has smaller number of groups with bigger number of organisms in each group. The latter partition i.e. P4 has bigger number of groups but with smaller number of organisms in each group. Some of the organisms were eliminated along the analysis to focus on a smaller group of organisms with higher similarity to the studied sequence.

The members in each of the group were aligned by partition to determine the conserved and neutral residues in the group. For each partition, the results from the group alignment will be compared and a consensus sequence was identified. The conserved residues are being translated on the trace record as they are, the class-specific residues are translated as X and highlighted in green while the neutral residues are shown as dashes (-).

The conserved residues identified from the partitions P0 and P4 respectively were being compared with the evolutionary trace analysis conserved residues, molecules from the Rasmol analysis and active site residues from PDB SUM. In partition P0, 109 residues out of the 376 residues or 29.8% residues are fully conserved throughout the whole group of 49 species. The conserved residues are mostly clustered in big groups, some are scattered in smaller groups and very few are scattered as single residues. 11 out of 23 active site residues identified using PDB SUM was fully conserved from the evolutionary trace analysis.

The protein mapping mapped conserved residues to the surface of the protein. The protein residues that are critical for structure and function were expected to be conserved throughout evolution. The conserved amino acid residues made up clusters mostly around the binding site.