Identification and Analysis of Novel Repeats and Domains in Bacterial, Archaeal and Human Proteomes

A Thesis

Submitted for the Degree of

DOCTOR OF PHILOSOPHY

By

G. R. HEMA LATHA

SCHOOL OF CHEMISTRY
UNIVERSITY OF HYDERABAD
HYDERABAD 500 046
INDIA

June 2008

Dedicated to...

My Beloved Mother and Father

CONTENTS

School of Chemistry
University of Hyderabad
Central University P. O.
Hyderabad 500 046
India

Statement

I hereby declare that the matter embodied in this thesis is the result of investigations carried out by me in the School of Chemistry, University of Hyderabad, Hyderabad, under the supervision of **Dr. Lalitha Guruprasad**.

In keeping with the general practice of reporting scientific observations, due acknowledgement has been made wherever the work described is based on the findings of other investigators.

G. R. HEMA LATHA

School of Chemistry
University of Hyderabad
Central University P. O.
Hyderabad 500 046
India

Certificate

Certified that the work embodied in this thesis entitled **"Identification and Analysis of Novel Repeats and Domains in Bacterial, Archaeal and Human Proteomes"** has been carried out by Ms. G. R. HEMA LATHA, under my supervision and the same has not been submitted elsewhere for a Degree.

Dr. LALITHA GURUPRASAD
(THESIS SUPERVISOR)

DEAN

SCHOOL OF CHEMISTRY

Acknowledgements

It gives me immense pleasure to express my deep sense of gratitude and profound respect to my teacher Dr. Lalitha Guruprasad for her inspiring guidance, valuable advice and personal motivation. She has been always helpful, approachable and extremely patient through out my tenure for which I am grateful to her.

I would like to acknowledge the suggestions and help of Dr. Guruprasad, CCMB.

I am thankful to Prof. E. D. Jemmis, Prof. M. Periasamy, and Prof. D. Basavaiah, Deans, School of Chemistry during my stay here, for providing the necessary facilities to carry out my research work. I thank all the faculty members of the School of Chemistry for their help and inspiring teaching.

I thank the funding bodies of CMSD for providing excellent computational facilities in the University. I thank Directors of CCMB and IICT for allowing me to use some of their computational and library facilities.

I thank Dr. Samar K. Das and Dr. Abani K. Bhuyan for their concern on various occasions.

I would like to acknowledge the suggestions and help of Dr Prof. Reddanna.

I would like to thank my colleagues Dr. Swathi, Krishna Kishore Inampudi, Srinivas, Karunakar and project students Nethu Singh, Satyanarayana, Manoj Kumar for creating a pleasant working atmosphere in the lab.

I would like to thank my project juniors Vijay, Sudheer, LSrinu, Chanu, Sankaraih, Sathish, Srinu (Chiru), Narayan, Om Narayan, Shaahid, Prabhat, Abirami and Nicee Srivastavaa.

All the non-teaching staff of the School is acknowledged for their help. I also thank Vinod Kumar, Rajender Reddy and other CMSD staff for their cooperation.

I express my warmest heartfelt thanks to my chinamma Aruna, Naga Vyjayanthi and B. Saritha, my uncle Raju and Anuradha aunty for their emotional support and timely help.

I express my warmest heartfelt thanks to my dearest friends Mrs. Latha and Mrs. Vijaya lakshmi (vijji) who deserve special mention for their endless affection, encouragement and continued moral support.

I would like to thank my M.Sc friends Upender, Ramesh, Shiva Prasad Vemula, Sudheer, Sunil, Nageshwar, Subhash and Sureka.

I would like to thank my friends Raghu, Santosh, Shankar, Anil (Tammulam Garu), Armugam, Nagaraju, Bhargavi, Biju, Raji, Vijay S, Rupa, Sri Latha, Anjali, Charu Sheela, Jyothi and Susheela.

I also acknowledge Dr. Mukkanti, Dr. Shivaiah, Dr. Supriya, Dr. Aparna, Dr. Yadaiah and Bhuvan for their healthy company through out my tenure.

I would like to thank Moin uncle, Azam uncle and Neerja madam for their timely help.

My special thanks are to my chinnana G. Venkatesh for his superior principles and moral support.

I also thank my friends Aparoy, VB Reddy, Dinakar, Padma and Usha.

All the research scholars of the School of Chemistry have been helpful and I thank them all.

I also thank my cousins Srikanth, Saikanth, Shivakanth and Renuka for their love and well wishes.

I am wordless to express my gratitude to my Mother and Father who have been the source of encouragement and guiding spirit and for their unfailing love all these years. I owe a great deal of loving thanks to my brother (Surender) and sister (Vanaja) for their support through out my career.

Financial support from the UGC New Delhi is greatly acknowledged.

G. R. Hema Latha

Abbreviations

PDB	:	Protein Data Bank
3-D	:	Three-dimensional
TMM	:	6-trehalose monomycolate
TDM	:	6, 6'-trehalose dimycolate
PDB-BLAST	:	Blast search against protein data bank
RMSD	:	Root Mean Square Deviation
PFAM	:	Protein families
DNA	:	Deoxyribo Nucleic Acid
RNA	:	Ribo Nucleic Acid
NMR	:	Nuclear Magnetic Resonance
PIR	:	Protein Information Resource
α	:	Alfa
β	:	Beta
TB	:	Tuberculosis
HSP	:	High Segment Pair
E	:	Expectation value
P	:	Probability score
HSP	:	High Scoring Pairs
Å	:	Ångström
X	:	any amino acid residue
IDs	:	Identities
ET	:	Evolutionary Trace
ETC	:	Evolutionary Time Cut-off
HGT	:	Horizontal Gene Transfer
ACR	:	Ancient Conserved Region
G+C	:	Guanine and Cytosine
E	:	Strand

H	:	Helix
MDP1	:	Mycobacterial DNA-binding protein1
PKD	:	Polycystic Kidney Domain
SLH	:	Surface Layer Homology
LRR	:	Leucine Rich Repeats
TPR	:	Tetratrico Peptide Repeats
SH2	:	Src homology 2
SH3	:	Src homology 3
PH	:	Pleckstrin homology
BLAST	:	Basic Local Alignment Search Tool
SMART	:	Simple Modular Architecture Research Tool
TRUST	:	Tracking Repeats Using Significance and transitivity
RADAR	:	Rapid Automatic Detection and Alignment of Repeats
REP	:	Repeat Finding method

Abstract

This thesis describes **Identification and Analysis of Novel Repeats and Domains in Bacterial, Archaeal and Human Proteomes.** It consists of six chapters 1) Introduction to bioinformatics tools in genomic data analysis, 3-D protein modeling and docking, 2) Analysis, 3-D structure modeling, docking and gene cluster identification of CMN mycolyl-transferases, 3) *In silico* method for the automated identification of novel repeats in complete proteomes, 4) Identification and analysis of novel amino acid sequence repeats and domains in *Bacillus anthracis* str. Ames proteome, 5) Identification and analysis of novel amino acid sequence repeats and domains in representative archaeal proteomes, and 6) Identification and analysis of novel amino acid sequence repeats and domains in human proteome. The work described in this thesis is exploratory in nature and is arranged in the order the investigations were carried out. Except the first chapter, all chapters are divided into Introduction, Methods, Results and Discussion, Conclusions, followed by References.

In the first chapter, a brief overview of the tools used in bioinformatics to characterize the protein sequences resulting from the genome sequencing projects is provided. Some commonly used tools such as BLASTP, PSI-BLAST, CLUSTALW and PHD are described. The databases such as SMART, PFAM, PROSITE and INTERPRO are also discussed. Various methods used to carry out the repeat identification such as RADAR, REP program, REPRO, PROSPERO and TRUST are described. A brief overview of evolutionary trace analysis, fold prediction, comparative structure modeling and an introduction to docking methods is provided.

The second chapter deals with the analysis, 3-D structure modeling, docking and gene cluster identification of CMN mycolyl-transferases. Tuberculosis (TB) is an infection caused by the bacterium *Mycobacterium*

tuberculosis. In *M. tuberculosis*, a major secreted protein complex antigen 85, constitutes three proteins antigen 85A, 85B and 85C that are responsible for the synthesis of cell envelope. These enzymes catalyze the transfer of mycolyl residue from one molecule of α, α'-TMM (trehalose monomycolate) to another TMM leading to the formation α, α' TDM (trehalose dimycolate) and are hence termed mycolyl-transferases and specifically present in CMN (Corynebacterium, Mycobacterium and Nocardia) genera. Mycolic acids are high molecular weight α- alkyl, β-hydroxy fatty acids that form a part of the unique cell envelope. The mycolic acids are named according to the individual genus from which they are isolated; in mycobacteria these are called eumycolic acids and possess long alkyl chain of length C_{60}-C_{90}, in nocardia these are called nocardiomycolic acids and possess short alkyl chain of length C_{40}–C_{50}, whereas in corynebacteria these are called corynemycolic acids and possess shorter alkyl chain of length C_{22}–C_{36}.

A comparative study of these proteins will be helpful in understanding their specificities and essential roles. We have carried out sequence similarity searches and identified proteins from mycobacteria, corynebacteria and nocardia. Multiple sequence analysis of the 31 mycolyl-transferases revealed that the 16 amino acid residues (L39, W51, P71, D81, W82, W97, F100, G124, S126, S150, D192, G214, E230, G260, H262 and W264) are conserved in all the sequences. We observed that the proteins of corynemycolyl-transferases and nocardiomycolyl-transferases have an insertion sequence of variable length (between 2 and 19 amino acid residues) and two proteins from *N. farcinica*, Nfa1810 and Nfa1820 consist of a 27 amino acid residue long insertion sequence rich in glycine and serine.

3-D models were constructed for these proteins using the crystal structure of Ag85B (PDB ID: 1F0P) as template structure. The 3-D models of corynemycolyl-transferases and nocardiomycolyl-transferases were compared with the crystal structure of mycolyl-transferases. Based on the structural

superposition, we observed major differences in the loop regions. The two proteins Nfa1810 and Nfa1820 consist of a insertion loop that is away from the active site. We observed that the proteins Nfa25110, Nfa45560, Nfa7210, Nfa38260, Nfa32420, Nfa23770, Nfa43800, Nfa30260, Dip0365, Ncgl0987, Ce1488, Ncgl0885, Ce0984, Ncgl2101, Ncgl0336, and Ce0356 accommodate insertion loops close to the substrate binding pocket. These proteins have large substrate binding pocket and mutations in amino acid residues that comprise substrate binding pocket and therefore, we propose that these enzymes may not bind trehalose. Some proteins were associated with variation in disulphide connectivity inspite of conservation in overall fold.

Based on gene cluster analysis, we have identified that the genes between Rv3799–Rv3808 in *M. tuberculosis* have orthologs in Corynebacteria, Mycobacteria and Nocardia (CMN) genomes. Therefore, this gene cluster possibly corresponds to the 'Ancient Conserved Region' of CMN mycolyl-transferases. The evolutionary trace analysis suggested that 12 amino acid residues: L39, W51, P71, W82, W97, F100, G124, S126, D192, E230, G260 and W264 are 'absolutely conserved'. These amino acid residues constitute the active site and conserved hydrophobic tunnel in CMN mycolyl-transferases. We observed the LGFP tandem repeats are also present in the C-terminal region of *N. farcinica* (Nfa1840) and *C. diphtheria* (Dip2193) proteins which imply that these are also functional cell surface proteins and may be involved in maintaining the cell wall integrity.

The third chapter deals with the *in silico* method for the automated identification of novel repeats in complete proteomes. The genes that code for proteins of unknown function are annotated as "Hypothetical proteins". However, more than 50% of proteins in the proteome zone remain unannotated and unidentified for function. The identification of repeats and domains in proteins is one such approach which can better explain the functions for

unannotated proteins or hypothetical proteins in the form of novel repeats and novel domains. A "repeat" corresponds to a region comprising less than 55 amino acid residues that occurs more than once, sometimes in tandem, along the protein primary sequence. A "domain" refers to a region of the protein comprising greater than 55 amino acid residues and does not contain internal sequence repeats. A repeat or domain type is characterized by specific conserved sequence motifs. Several web-based methods are available for *ab initio* identification of sequence repeats in proteins. The popular programs that identify internal repeats in proteins are RADAR, REP Program, REPRO, PROSPERO and TRUST.

We have used TRUST as the main program for novel repeat identification method in complete proteomes of bacterial, archaeal and human genomes. TRUST program exploits the concept of transitivity of alignments as well as a statistical scheme optimized for the evaluation of repeat significance. It detects repeats using the Waterman-Eggert algorithm. Starting from significant local sub-optimal alignments, the application of transitivity allows to: 1) identify distant repeat homologs for which no alignments were found; 2) gain confidence about consistently well aligned regions; and 3) recognize and reduce the contribution of non-homologous repeats. This assessment step will enable to derive a virtually noise free profile representing a generalized repeat with high fidelity. TRUST is a useful and reliable tool for mining tandem and non-tandem repeats in protein sequence databases, to predict multiple repeat types with varying intervening segments within a single sequence.

We have downloaded individual proteomes for example bacterial (Ex. *Bacillus anthracis* str. Ames), archaeal (13 representative organisms from archaeal origin) and human proteome from the NCBI website in FASTA format.

We have downloaded and installed TRUST on the local Pentium IV computers on the Linux platform. Linux shell scripts were written to automate TRUST and subsequent steps for repeat identification. The details of this

method are provided in Chapter 3. The TRUST program was run for all the sequences in each proteome. We can submit up to 5 organisms for repeat identification using TRUST as long as the total size of the file does not exceed 2MB. Based on the size of the TRUST output file, the protein sequences with no internal repeats were discarded automatically. i.e., only those protein sequences which comprise repeats were retained. Thus selected proteins were submitted to SMART online program in batch mode. Manual inspections of the SMART results identified proteins comprising known repeats or domains and were therefore discarded. Only those repeats that were not identified by SMART database were retained for further analysis. Using automatic shell scripts, these protein sequences were then analyzed using offline PSI-BLAST program for three iterations against the NCBI NR database and WU-BLAST2 against UNIPROT database. The proteins confirmed to comprise repeats by the BLAST program were retained and were tested for presence in the offline versions of INTERPRO (*Database: iprscan_DATA_10.0, Applications: iprscan_V4.1, iprscan_binn4.x_Linux*) and PFAM (*release date: April 26, 2005*) databases. A final check was made using online versions of INTERPRO and PFAM.

The repeats which are not present in any of these databases were considered to be novel repeats or domains, depending upon (1) the number of times they occur in the protein sequences and (2) length of the amino acid sequence region. The novel repeats and domains thus identified were subjected to online PSI-BLAST analysis in order to identify other proteins from databases that comprise these repeats and domains. Multiple sequence alignment program, CLUSTALW was used to detect the extent of sequence conservation and the secondary structure prediction was carried out using PHD and PSIPRED methods. The programs developed in this work save a large amount of time and labor involved in similar studies.

The fourth chapter deals with the *in silico* identification and analysis of novel amino acid sequence repeats in *Bacillus anthracis* str. Ames proteome. The anthrax is a disease of herbivores and other mammals including humans, caused by the *B. anthracis* str. Ames, a Gram-positive, rod-shaped, non-motile, spore forming bacterium. It is an endospore forming bacterium that causes inhalational anthrax. Expression of the major plasmid encoded virulence determinants, tripartite toxin and a poly-D-glutamic acid capsule are essential for full pathogenicity. Key virulence genes found on plasmids are pXO1 and pXO2. The 60 MDa plasmid pXO2 carries genes required for the synthesis of an antiphagocytic poly-D-glutamic acid capsule. The 110 MDa plasmid pXO1 is required for the synthesis of the anthrax proteins, edema factor, lethal factor and protective antigen. The complete genome sequence of *B. anthracis* str. Ames is available and it comprises of 5,227,293 base pairs and 5,508 genes with an overall G+C content of 35.4%. Of these, 2,762 are functional genes, 1,212 are conserved hypothetical genes, 657 genes are of unknown function and 877 genes are annotated as hypothetical proteins.

In this work, we have systematically identified and analyzed 4 repeats and 10 domains using TRUST. These correspond to: 1) 57 amino acid residue PxV domain, 2) 122 amino acid residue FxF domain, 3) 111 amino acid residue YEFF domain, 4) 109 amino acid residue IMxxH domain, 5) 103 amino acid residue VxxT domain, 6) 84 amino acid residue ExW domain, 7) 104 amino acid residue NTGFIG domain, 8) 36 amino acid residue NxGK repeat, 9) 95 amino acid residue VYV domain, 10) 75 amino acid residue KEWE domain, 11) 59 amino acid residue AFL domain, 12) 53 amino acid residue RIDVK repeat, 13) a) 41 amino acid residue AGQF repeat and b) 42 amino acid residue GSAL repeat. We have predicted the secondary structures for these repeats and domains. Some of them were found to be associated with specific functions. For example, the NxGK repeats are associated with SAP domain. The SAP domain is a DNA-binding motif that is involved in chromosomal organization.

Therefore, we believe that these repeats also participate in a similar function. The YEFF domain containing proteins are associated with RGD motif and may be involved in cell adhesion. The RIDVK, AGQF and GSAL repeats are specifically present only in *B. anthracis* str. Ames and are orphan proteins. From the presence of VYV and AFL domains in all the *B. anthracis* species and their absence in *B. cereus* genomes, we identified the differences between these two genomes that are otherwise closely related. The identification of novel repeats and domains corresponding to *B. anthracis* str. Ames proteome may be useful for annotation.

The fifth chapter deals with *in silico* identification and analysis of novel amino acid sequence repeats and domains in representative archaeal proteomes. Archaea is a major division of living organisms. Archaea are distinguished from other organisms by three major criteria: 1. their 16S rRNA sequences are different from those of eubacteria and eukaryotes, 2. their cell walls consist of glycosylated proteins rather than peptidoglycan structure in eubacteria and 3. their membrane lipids are unique, consisting entirely of derivatives of an ether linked isoprenoid structure.

Phylogenetic analysis of small-subunit rRNA sequences distinguishes two distinct archaeal sub-domains: the euryarchaeotes and the crenarchaeotes. The euryarchaeotes include methanogens, halophiles, and sulfur reducing thermophiles. The euryarchaeota is further divided into nine families. They are as follows 1. Archaeoglobales, 2. Halobacteriales, 3. Methanobacteriales, 4. Methanococcales, 5. Methanopyrales, 6. Methanosarcinales, 7. Thermococcales, 8. Thermoplasmales and 9. Thermoplasmatales .

The crenarchaeotes share a 16S rRNA signature with the euryarchaeotes within the archaeal domain. The crenarchaeotes are in many instances sulfur dependent thermophiles and have initially been regarded as more homogenous than the euryarchaotes. The crenarchaeota is further divided into three families.

They are 1. Desulfurococcales, 2. Sulfolobales and 3. Thermoproteales. Nanoarchaeota is the newly identified domain of archaea and *Nanoarchaeum equitans* belongs to this domain.

The complete and nearly complete sequencing of archaeal genomes will provide data to infer properties of proteins that must have been present in a common ancestor, as well as properties that may pinpoint the basis of divergence. Since many proteins in these genomes are identified from genome sequencing projects, they are hypothetical and yet to be characterized. In order to further characterize these hypothetical proteins we have carried out a systematic identification and analysis of the novel amino acid sequence repeats of all the available representative archaeal proteomes using computational tools.

We have identified and analyzed 56 domains and 38 repeats in 13 archaeal proteomes according to the representative phylogeny. These repeats and domains have not been reported before in archaeal proteomes and are novel. They are as follows: 1. *Aeropyrum pernix* K1 (1 domain), 2. *Sulfolobus tokodaii* str. 7 (7 domains and 5 repeats), 3. *Pyrobaculum aerophilum* str. IM2 (5 domains and 4 repeats), 4. *Archaeoglobus fulgidus* DSM 4304 (7 domains and 4 repeats), 5. *Halobacterium salinarium* NRC-1 (8 domains and 1 repeat), 6. *Methanobacterium thermoautotrophicum* str. Delta H (4 domains and 2 repeats), 7. *Methanocaldococcus jannaschii* DSM 2661 (5 domains and 2 repeats), 8. *Methanopyrus kandleri* AV19 (2 domains), 9. *Methanosarcina acetivorans* str. C2A (8 domains and 13 repeats), 10. *Pyrococcus abyssi* GE5 (4 domains), 11. *Thermoplasma acidophilum* DSM 1728 (4 domains), 12. *Picrophilus torridus* DSM 9790 (6 repeats), 13. *Nanoarchaeum equitans* Kin4-M (1 domain and 1 repeat).

We observed that the 100 amino acid residues GQP domain of *S. tokodaii* str. 7 belongs to COG1449 (Cluster of Orthologues) and the domain is predicted to function as sugar transporter permease protein. The 108 amino acid residues NDFA domain of *P. aerophilum* str. IM2 belongs to COG0438M

and is predicted to function as trehalose-6-phosphate synthase. The 83 amino acid residues CCE domain of *A. fulgidus* DSM 4304 has been described as a cell surface protein and we propose that these are cell surface protein specific repeats. We have predicted the functions to some other novel repeats and domains from *H. salinarium* NRC-1, *M. kandleri* AV19, *M. acetivorans* str. C2A, *P. abyssi* GE5, *T. acidophilum* DSM and *P. torridus* DSM 9790. From the repeats and domains present in these representative archaeal genomes (the data previously known and the findings in this work), we infer that *N. equitans* Kin4-M is a minimalist archaea. The exchange of genes between archaeal and bacterial genomes is maximal in *M. acetivorans* C2A and has therefore undergone extensive evolution. The number of orphan proteins comprising repeats and domains is also high, indicating a significant variation in evolution of these genomes. This is required for the adaptation of individual organisms to extreme living conditions such as high temperature, pressure and pH.

The sixth chapter deals with *in silico* identification and analysis of novel amino acid sequence repeats and domains in human proteome. The Human Genome Project (HGP) was launched in 1990 with the goal of obtaining a highly accurate sequence of the vast majority of the euchromatic portion of the human genome. A predominant part of the human genome consists of repetitive sequences of various types encompassing large segmental duplications, interspersed transposon derived repeats and tandem repeats. Amino acid tandem repeats, also known as homopolymeric tracts, is a very common feature of eukaryotic proteins. They are present in nearly one-fifth of human gene products. Human proteins contain more amino acid repeats than rodent proteins and the trinucleotide repeats are also more abundant in human coding sequences. The uncontrolled expansion of trinucleotide repeats within human coding sequences is associated with several neurodegenerative disorders. Examples are Huntington's disease and dentatorubropallidolusyan atrophy, both

associated with abnormally long expansions of CAG runs encoding poly-glutamine tracts.

Repeat structures in proteins have recently been found to play vital roles in various biological functions ranging from signal transduction, transcription regulation, to apoptosis, and are also recognized by their association with several human diseases. It is of paramount importance to identify the structures of the individual protein repeats lying within the human proteome and explore their protein interaction mechanisms to understand the complex biological processes and the human body in itself. Realizing the importance of amino acid repeats in the proteome and in human disorders, we undertook a study to identify and analyze the novel amino acid sequence repeats that are not present in any of the known databases and that are not reported so far with the available draft sequence of human genome.

In this work, we have identified 7 domains and 18 repeats using TRUST. The domains are as follows: 1. 58 amino acid residue GPA domain, 2. 61 amino acid residue RxH domain, 3. 68 amino acid residue GLG domain, 4. 71 amino acid residue SAS domain, 5. 73 amino acid residue WKRK domain, 6. 85 amino acid residue FSS domain and 7. 109 amino acid residue LLE domain. The RxH, WKRK, FSS and LLE domains are present in *Homo sapiens* and other eukaryotic genomes, GPA domain is present in *Homo sapiens* and *Pan troglodytes* genomes. The GLG and SAS domains are *Homo sapiens* specific and are orphan proteins.

The repeats are as follows: 1. 30 amino acid residue PGQY repeat, 2. 31 amino acid residue FYE repeat, 3. 34 amino acid residue VHMM repeat, 4. 34 amino acid residue TQG repeat, 5. 51 amino acid residue PES repeat, 6. 34 amino acid residue HTQ repeat, 7. 38 amino acid residue PTT repeat, 8. 34 amino acid residue FSQ repeat, 9. 36 amino acid residue PEG repeat, 10. 42 amino acid residue SSC repeat, 11. 42 amino acid residue YCL repeat, 12. 43

amino acid residue VSR repeat, 13. 54 amino acid residue ALPG repeat, 14. 43 amino acid residue SVT repeat, 15. 49 amino acid residue CDxD repeat, 16. 50 amino acid residue GGF repeat, 17. 52 amino acid residue NYS repeat and 18. 52 amino acid residue RPE repeat. The PGQY, FYE, VHMM, TQG, PES, FSQ, SSC, YCL, VSR, ALPG, CDxD and RPE repeats are present in *Homo sapiens* and other eukaryotic genomes while the HTQ, PTT, PEG, SVT, GGF and NYS repeats are *Homo sapiens* specific and are orphan proteins.

Many of the domains and repeats identified were observed to be associated with disease causing proteins. The 61 amino acid residue RxH domain encodes PDZ domain containing proteins which play prominent roles in synapse formation and we predict a similar function for the RxH domain. The 73 amino acid residue WKRK domain is associated with Williams-Beuren syndrome (WBS; OMIM 194050), that is caused by heterozygous deletions of ~1.6 Mb of chromosomal sub-band 7q11.23. The 34 amino acid residue HTQ and 38 amino acid residue PTT repeats encodes the Polycystic kidney disease 1 like 3 proteins. Polycystic kidney disease (PKD) is a disease of the nephron, characterized by the formation of multiple renal tubular cysts, leading to endstage renal failure and therefore, we predict a similar function for the HTQ and PTT repeats. Further database searches identified that some novel repeats and domains are also present in other mammalian genomes. Thus, the identified novel repeats and domains of human proteome can be used for annotation in the databases.

CHAPTER 1

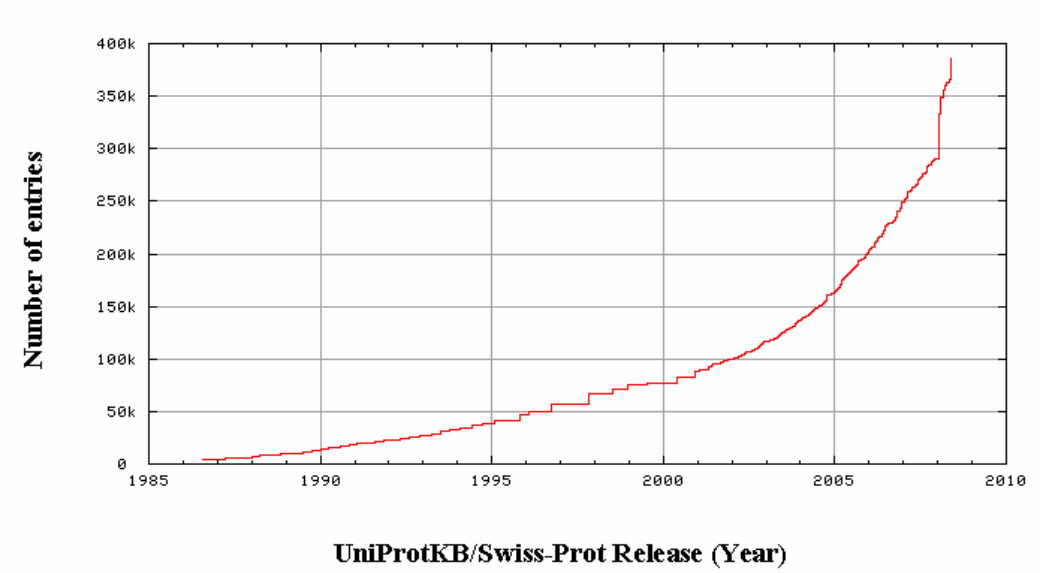
Introduction to Bioinformatics Tools in Genomic Data Analysis, 3-D Protein Modeling and Docking

1.1 Bioinformatics tools in genomic data analysis

The recent flow of data from genomics has given rise to a new field called "Bioinformatics", which is a combination of biology and informatics. Its objective is to understand and organize biological information on a large-scale. The available information which needs to be analyzed falls within different categories in biology such as (i) genome, DNA and protein sequences, (ii) DNA and protein structures, (iii) RNA and protein expression data, (iv) molecular interactions, for instance, between proteins, (vi) physiological high-throughput data of metabolites or protein and (vii) the interplay between structure and function in the evolution of diverse biological systems, (viii) the components of biological systems and literature (Hocquette, 2005). The objectives of bioinformatics are therefore different depending upon the initial dataset.

The development of bioinformatics has however taken biology away from the wet laboratory to some extent. Bioinformatics is now a branch of science *per se*. We must also be aware that, with the sequencing of many genomes, the scientific questions have shifted from identifying genes to discovering their functions. The bioinformatics tools are continuously changing to adapt themselves to the new queries of knowledge retrieval (Fraser & Marcotte, 2004). As more and more systems biology approaches are used to investigate the different types of biological macromolecules, increasing numbers of whole genomic studies are now available for a large array of organisms. Whether it is genomics, transcriptomics, proteomics, interactomics or metabolomics, the full complement of genomic information at different levels can be juxtaposed between different organisms to reveal similarities or differences and even to provide consensus models. At the intersection of comparative genomics and systems biology lies great possibility for discovery, analysis and prediction (Lin & Qian, 2007).

Figure 1.1: Graph showing the explosion of sequence data (as on 10-06-2008) in UniProtKB/SWISS-PROT Database (taken from http://www.expasy.ch/ sprot/relnotes/ relstat.html).



The wealth of sequence information brought about by the genome sequencing projects has led to the discovery of several computational tools, which enables the researchers to analyze the genes and proteins in whole genomes. These computational methods have been developed to solve the biological problems, using DNA, protein sequences and other related information.

The information from the completed genome sequence projects are stored in the databanks such as EMBL, GenBank for DNA and SWISS-PROT, UniProt, NRDB for proteins and are freely available to the public. Millions of sequences are available in these databanks that provide basic information about the respective entries. Graph shown in the Figure 1.1 represents the explosion of information content in the UniProtKB/SWISS-PROT database.

Though DNA is the genetic material, it does not carry out the processes of life. This genetic code is transcribed and translated in the synthesis of protein molecules, which are present as the structures and molecular machines that

make the cell function. Proteins contribute to almost all events in the cells of a living organism. The polypeptide chain of a protein folds into a specific 3-D structure, which governs its function. Several developments in the techniques of structure determination at atomic resolution, X-ray diffraction and nuclear magnetic resonance (NMR) spectroscopy, have enhanced the quality and speed of structural studies (Zhang & Kim, 2003). Nevertheless, current statistics still show that the known protein sequences vastly outnumber the available protein structures (51,366) deposited in protein data bank (PDB) so far. This is due to the inability to express, purify and crystallize some proteins as well as the intrinsic limitations of the structure determination techniques.

It becomes a challenge for the researchers to annotate this huge genomic data. About 40-50% of proteins in each genome are novel and are not biochemically and structurally characterized. Experimental characterization of each sequence is however time consuming. Therefore, adding value to the structure and function of these novel proteins by means of comparative studies, using computational tools is one of the challenges to the researchers worldwide. Sophisticated mathematical, statistical and computational techniques are developed to handle, analyze and add value to this flood of data. These studies have now become one of the frontier areas of research in the modern biology.

Brief descriptions of some protein sequence analysis tools, 3-D structure modeling and docking approaches with emphasis on the methods and programs that have been used to carry out the work embodied in this thesis are discussed in this chapter.

1.1.1 Sequence analysis tools: As the volume of genome sequence data now available is enormous with more than 750 genomes being either completely sequenced or in progress, a biologist is using several databases with increasing attention towards finding any novel "Genes" or "Proteins" or "Functions". However, various analysis based on sequence, structure, function and "Omic" data have revealed consensus in annotation of different sets of predicted genes.

1.1.1.1 Database searching:

DNA and protein databases: DNA sequence databases were first assembled at Los Alamos National Laboratory (LANL), by Walter Goad and colleagues in the GenBank database and at the European Molecular Biology Laboratory (EMBL) in Heidelberg, Germany. GenBank is now under the auspices of the National Center for Biotechnology Information (NCBI) (http://ncbi.nlm.nih.gov). The EMBL Data Library was founded in 1980 (http://www.ebi.ac.uk). The EMBL maintains DNA and protein sequence databases. In 1984 the DNA DataBank of Japan (DDBJ) came into existence (http://ddbj.nig.ac.jp). GenBank, EMBL and DDBJ have now formed the International Nucleotide Sequence Database Collaboration (http://www.ncbi.nlm.nih.gov/collab), which acts to facilitate exchange of data on a daily basis. Translated nucleotide sequence information is included in the Protein Information Resource (PIR) database at the National Biomedical Research Foundation in Washington, DC. GenBank(R) is a comprehensive database of publicly available DNA sequences for more than 205,000 named organisms and for more than 60,000 within the embryophyta, obtained through submissions from individual laboratories and batch submissions from large-scale genome sequencing projects. GenBank is accessible through the NCBI retrieval system, Entrez, which integrates data from the major DNA and protein sequence databases with taxonomy, genome mapping, protein structure and domain information. SWISS-PROT is a curated protein sequence database which strives to provide a high level of annotation

(such as the description of the function of a protein, its domains structure, post-translational modifications, variants, etc.), with minimal redundancy and high level of integration with other databases (Bairoch & Apweiler, 2000). The Universal Protein Resource (UniProt) provides a stable, comprehensive, freely accessible, central resource of protein sequences and functional annotation. The UniProt Consortium is collaboration between the European Bioinformatics Institute (EBI), the PIR and the Swiss Institute of Bioinformatics (SIB). Their core activities include manual curation of protein sequences assisted by computational analysis, sequence archiving, development of a user friendly UniProt website and the provision of additional value-added information through cross references to other databases (The Uniprot Consortium, 2008).

The "nr" (non-redundant) database is the largest nucleotide database available through NCBI. It includes all GenBank, RefSeq Nucleotides, EMBL, DDBJ and PDB sequences. The format of a database entry is such that each sequence file contains the information about the assigned accession number, source organism, function of the sequence, literature references, location of mRNAs, coding regions, positions of important mutations and sequence.

Comparison of a sequence with entries in a database is required to identify similar sequences that share homology. This can be done at both nucleotide and protein level. After proper validation of the results, multiple sequence alignments of these related sequences can be built using consensus sequences of protein families that help in the identification of domains, motifs or functional sites. Detection of sequence similarity among different proteins has led to the classification of proteins on the basis of structure and function. It has been observed that most often similar sequences share similar structure and function. In addition, database searches are also used as primary requirement in identifying a structural homolog of a protein sequence. The most widely used programs for database searching are BLAST (Altschul *et al.,* 1990) and FASTA (Pearson & Lipman, 1988, http://www.ebi.ac.uk/fasta33/).

Basic **L**ocal **A**lignment **S**earch **T**ool **(BLAST):** The BLAST program is used to identify sequence similar homologs from DNA or protein databases. The program takes a query sequence and searches it against the database selected by the user. It aligns the query sequence against every subject sequence in the database and the results are reported in the form of a ranked list followed by a series of individual sequence alignments, plus various statistics and score parameters. Every hit in the list is assigned with a similarity score S. Further, this score is analyzed to calculate the extent of such matching to occur by chance. For this purpose E-value is calculated for every hit. BLAST program finds regions of local similarity and calculates the statistical significance of matches (Altschul *et al.,* 1990) (http://www.ebi.ac.uk/blast2) and (http://www.ncbi.nlm.nih.gov/blast/Blast. cgi/).

The BLAST program first dissects the query sequence into words of length k (3 for proteins and 11 for nucleotides). These words are searched against the database for matches and scores are assigned with either BLOSUM (Henikoff & Henikoff, 1992) or PAM (Dayhoff, 1978) scoring matrices. Word hits that score more than T (neighborhood word score threshold) are extended in both directions to generate an alignment between segment pairs. The "T" parameter dictates the speed and sensitivity of the search. The extension process is stopped when the scores drop from its maximum achieved score and the segment pairs are referred as high scoring pairs (HSP). The next step is to determine those HSPs of sequences, which have score greater than a cut-off score (S). S is determined empirically by examining a range of scores found by comparing random sequences and by choosing a value that is significantly greater. BLAST determines the statistical significance of HSPs and generates sequence hits in the descending order of E (expectation value) and P (probability score) values. E and P values are different ways of representing the significance of the alignment. The highly significant E or P values will be those close to 0 and lower values. BLAST also filters the low-complexity regions.

Filtering is done by SEG and XNU filters and applied to the query sequence alone to make the search focus on more important parts of the sequence. These regions are marked with X in protein sequences and N in nucleotide sequences and are then ignored by BLAST.

The BLASTP offers various user defined options. A choice can be made on database to be searched. Based on the requirement, a user can switch to PDB or SWISS-PROT database or a specific organism. Other options include selection of matrices, filters, adjustment of sensitivity and number of alignments etc. The default parameters for BLASTP include BLOSUM62 scoring matrix, a value of 11 is assigned for gap opening and a value of 1 for gap extension.

BLAST uses Smith-Waterman dynamic programming algorithm (Smith & Waterman, 1981a, 1981b). It detects local as well as global alignments using a heuristic approach. The exhaustive Smith-Waterman approach is too slow for searching large genomic databases such as GenBank. Therefore, the BLAST algorithm uses a heuristic approach that is slightly less accurate than Smith-Waterman, but that is over 50 times faster. There are five different BLAST programs, which can be distinguished by the type of the query sequence (DNA or protein) and the type of the subject database for searching.

BLASTP-compares an amino acid query sequence against a protein sequence database.

BLASTN-compares a nucleotide query sequence against a nucleotide sequence database.

BLASTX-compares the six-frame conceptual translation products of a nucleotide query sequence (both strands) against a protein sequence database.

TBLASTN-compares a protein query sequence against a nucleotide sequence database dynamically translated in all six reading frames (both strands).

TBLASTX-compares the six-frame translations of a nucleotide query sequence against the six-frame translations of a nucleotide sequence database.

Position **S**pecific **I**terative **BLAST (PSI-BLAST):** PSI-BLAST program is used for finding distant relatives of a protein. The program makes a list of all closely related proteins. These proteins are then combined into a profile that represents an average sequence. A query against the protein database is then run using this profile and a larger group of proteins are found. This larger group is used to construct another profile and the process is repeated (Altschul *et al.,* 1997) till one finds all related proteins in the database. This method is more reliable and used in several other programs such as PSIPRED and PHD that are secondary structure prediction methods. By including related proteins in the search, PSI-BLAST is much more sensitive in identifying distantly related proteins than using the standard protein-protein BLAST (http://www.ncbi.nlm.nih.gov/blast/Blast.cgi/).

Pattern **H**it **I**nitiated **BLAST (PHI-BLAST):** PHI-BLAST is a search program that combines matching of regular expressions with local alignments surrounding the match. The calculation of local alignments is done using a method very similar to gapped BLAST (Zhang, 1998). The most important features of the program have been incorporated into the BLAST framework partially for user convenience and partly so that PHI-BLAST may be combined seamlessly with PSI-BLAST. PHI-BLAST is the most preferred search tool for pattern occurrences because it filters out those cases where the pattern occurrence is probably random and not indicative of homology. PHI-BLAST may be preferable to other types of BLAST programs because it is faster and allows the user to express a rigid pattern occurrence requirement. PHI-BLAST uses Baeza-Yates and Gonnet (Baeza, 1992) and Wu and Manber, 1992 algorithm, which permits simple patterns to be represented in a single computer word and matches to be found very efficiently. PHI-BLAST was specifically designed to combine pattern search to find statistically significant sequence similarity (http://www.ncbi.nlm.nih.gov/blast/Blast.cgi/).

1.1.1.2 Motif/Pattern: A sequence motif is a short conserved region found in a number of related protein sequences. Motifs often correspond to core structural and functional elements of the proteins. Their conserved nature allows them to be used to diagnose family membership and predict function. Genome sequencing provides the basis for a systematic analysis of all motifs that are present in a particular organism. Protein sequences can be searched for the presence of known motifs in databases such as PROSITE (http://www.expasy.org/prosite/), ProDom (http://prodom.prabi.fr/prodom/current/html/form.php), PFAM (http://www.sanger.ac.uk/Software/Pfam/) and PRINTS (http://www.bioinf.manchester.ac.uk/dbbrowser/PRINTS). These pattern and profile searches constitute an important resource for the classification of majority of the newly appearing protein sequences into one of the known families.

PROSITE: PROSITE is a database of protein families and domains. It consists of a large collection of biologically meaningful signatures that are described as patterns or profiles. Each signature is linked to a documentation that provides useful biological information on the protein family, domain or functional site identified by the signature. More than 200 domains have been added to the PROSITE database over the past 2 years and 52% of UniProtKB/SWISS-PROT entries (release 48.1 of September 27, 2005) have a cross-reference to a PROSITE entry (Hulo *et al.,* 2006).

1.1.1.3 Protein families and domains: Protein families are the groups of molecules that share a significant sequence similarity and a common evolutionary history. Proteins within a family preserve their molecular structure and thus can maintain similar or even identical biochemical function across vast evolutionary distances. Many proteins are modular in nature. The modules are structural units or domains that are covalently linked to generate multi-domain proteins. Each Domain is a structural and functional unit that has a specific

biochemical activity. A brief description of protein domains is discussed in section 1.4.2.

SMART (Simple **M**odular **A**rchitecture **R**esearch **T**ool**):** The SMART is an online resource (http://smart.embl.de/) used for protein domain identification and the analysis of protein domain architectures (Letunic *et al.,* 2006). SMART offers a high level of sensitivity and specificity coupled with ease of use. It contains several unique aspects, including automatic seed alignment generation, detection of repeated motifs or domains and a protocol for combining domain predictions from homologous subfamilies. Visualization tools have been developed to allow analysis of gene intron-exon structure within the context of protein domain structure and to align these displays to provide schematic comparisons of orthologous genes or multiple transcripts from the same gene. It also allows batch retrieval of multiple entries.

INTERPRO: INTERPRO is an integrated database resource for protein families, domains and functional sites (Mulder *et al.,* 2005). INTERPRO provides integrated view of the commonly used protein signature databases such as PROSITE, PRINTS, ProDom, PFAM, SMART, TIGRFAMs, PIRSF, SUPERFAMILY, Gene3D and PANTHER. Signatures are manually integrated into INTERPRO entries that are curated to provide biological and functional information. It also provides links to additional reading fields, new database links, extensions to the web interface and additional match XML files. INTERPRO covers over 78% of all proteins in the UniProtKB database (Mulder *et al.,* 2007). The database is available for text and sequence based searches via the web server (http://www.ebi.ac.uk/InterProScan/).

PFAM: PFAM is a comprehensive collection of protein domains and families and helps in the genome annotation (Bateman *et al.,* 2004). Each family in PFAM is represented by multiple sequence alignments and Hidden Markov

Model (HMM) profile and can be used to view the domain organisation of proteins. Structural data has been utilised to ensure that families in PFAM correspond to structural domains and to improve domain based annotation. Predictions of non-domain regions are also included. In addition to secondary structure, PFAM multiple sequence alignments now contain active site residues highlighted. New search tools, including taxonomy search and domain query, add to the functionality and usability of the PFAM resource.

Apart from the well known annotated domains, PFAM also provides the information of functionally uncharacterized families, known as Domains of Unknown Function (DUFs) and Uncharacterized Protein Families (UPFs). DUFs are families that have been created by PFAM and UPFs are those created by SWISS-PROT and added to PFAM database. PFAM covers over 9,318 protein families. PFAM is now based not only on the UniProtKB sequence database, but also on NCBI GenPept and on sequences from selected metagenomics projects (Finn *et al.,* 2008). The database is available for text and sequence based searches via the web server (http://www.sanger.ac.uk/Software/Pfam).

1.2 Multiple sequence alignment: Multiple sequence alignment (MSA) is an important aspect of sequence analysis which is routinely used to identify and measure similarities between samples of DNA, RNA or protein. An alignment is the vertical arrangement of sequences of 'residues' (nucleotides or amino acids) that maximizes the similarities between them. The relationships between sequences are very complex since they have been exposed to evolutionary pressures and mutations over millions of years. A multiple sequence alignment arranges three or more sequences, such that residues with common structural positions and / or ancestral residues are aligned in the same column in a group of sequences and gaps are inserted in the sequences, if required. If two sequences in an alignment share a common ancestor, mismatches can be

interpreted as point mutations and gaps as insertion or deletion mutations, that are introduced in one or both lineages, in the time since they diverged from one another. In protein sequence alignment, the degree of similarity between amino acids occupying a particular position in the sequences can be interpreted as a rough measure of the conservation of a particular region or sequence motif lineage. The most similar regions in the multiple sequence alignment may represent structural domains or regions of functional importance.

Multiple sequence alignments often provide an understanding of evolutionary history of sequences. If the sequences in the alignment are very well conserved, then it implies that these sequences are recently derived from a common ancestor sequence. The function and structure of an unknown protein is predicted by aligning its sequence with others of known function and structure and also in the prediction of probes for the same family of sequences in the same or different organisms. Multiple sequence alignments can build consensus sequences of known families, domains, motifs or sites. Combining these predictions with primary biochemical data can provide valuable insights into protein structure and function.

CLUSTALW: CLUSTALW is a fully automated program for global multiple alignment of nucleotide and protein sequences. This is very useful in designing experiments to test the function of specific proteins, in predicting the function and structure of proteins and in identifying new members of protein families. CLUSTALW generates multiple sequence alignments and a phylogenetic tree.

CLUSTALW produces biologically meaningful multiple sequence alignments of divergent sequences. It calculates the best match for the selected sequences and aligns them up so that the identities, similarities and differences can be seen. The alignment in CLUSTALW is achieved via three steps: 1) pair-wise alignment, 2) guide-tree generation and 3) progressive alignment. Evolutionary relationships can be observed in a diagrammatic form by viewing

Cladograms or Phylograms. It can manipulate existing alignments and carry out profile analysis (Thompson *et al.,* 1994). The majority of the automated multiple sequence alignments are based on the progressive approach of the Feng and Doolittle (Feng & Doolittle, 1987). CLUSTALW, developed by Thompson *et al.,* 1994, incorporated a number of improvements to the alignment algorithm, including sequence weighting, position-specific gap penalties and the choice of a suitable residue comparison matrix at each stage in the multiple alignments.

In CLUSTALW alignment, scores can be calculated by two methods, slow / accurate or fast / approximate, that use dynamic programming (Smith & Waterman, 1981a; 1981b) and Wilbur and Lipman methods (Wilbur & Lipman, 1983) respectively. CLUSTALW provides several options, such as use of slow or fast pair-wise alignments, DNA or protein sequences, protein weight matrix, gap open, gap extension, end gaps and gap distances. The default parameters for protein sequences are: Protein Gap Extension Penalty = 0.2; Protein matrix = Gonnet; Protein ENDGAP = -1; Protein GAPDIST = 4. This program is available for sequence based searches via the web server (http://www.ebi.ac.uk/Tools/clustalw2/).

1.2.1 Evolutionary Trace analysis: Sequence conservation during evolution is the foundation for the functional classification of the enormous number of new protein sequences being discovered in the current era of genome sequencing. A crucial aspect in protein sequence analysis is the identification of functional sites such as ligand-binding sites, active sites, protein-protein interaction sites, signal sequences and post-translational modification sites. Traditionally, the residues that are conserved in all members of a protein family are assembled as motifs and correlated to the main function of that protein family. Given the massive increase in the number of new sequences and structures, a critical problem is to integrate these raw data into meaningful biological information.

Evolutionary Trace (ET) is a method that uses a sequence similarity tree of a family of homologous proteins to highlight residues, which are statistically likely to be under evolutionary pressure and therefore, of functional or structural importance for the family (Lichtarge *et al.,* 1996). When the structure is available, ET results may be mapped onto the structure, thus outlining known as well as putative functional parts of the protein surface. A trace is generated by comparing the consensus sequences for groups of proteins that originate from a common node in a phylogenetic tree and is characterized by a common evolutionary time cut-off (ETC) and classifying each residue as one of the three types: absolutely conserved, class-specific and conserved. Here, class specific denotes residues occupying a strictly conserved location in the sequence alignment, but differing in the nature of their conservation between various sub-groups. The information obtained by the ET method can then be mapped onto known protein structures, thus allowing us to identify clusters of important amino acid residues and to distinguish between buried and exposed residues. The strength of the ET method lies in its flexibility: depending on the ETC value for which a trace is generated, it is possible to maximize the specificity of the analysis over its sensitivity and vice versa. It allows for a wide range of functional resolution (Innis *et al.,* 2000).

Several servers are available to rank protein residues according to the estimated evolutionary pressure they experience. One such server is the "TraceSuite II" which uses phylogenetic information to rank the residues in a protein sequence by evolutionary importance and then map those ranked at the top onto a representative structure. If these residues form structural clusters, one can identify functional surfaces such as those involved in molecular recognition. For the ET analysis, TraceSuite II server is available at the website (http://www-cryst.bioc.cam.ac.uk/jiye/evoltrace/evoltrace.html).

1.3 Secondary structure prediction methods: Secondary structure is defined as the patterns of hydrogen bonds between backbone amide groups within proteins and consists of local inter-residue interactions mediated by hydrogen bonds in a protein. It is the spatial arrangement of three types of sub-structures known as helices, strands and coils in a protein. The most common secondary structures are alpha helices and beta sheets.

Alpha-helix is a right-handed coiled conformation, resembling a spring, in which every backbone of N-H group of amino acid (n) donates a hydrogen bond to the backbone C=O group of the amino acid four residues earlier (n + 4). Beta sheet consists of stretch of amino acids connected laterally by three or more hydrogen bonds, forming a twisted, pleated sheet. A coiled coil is a structural motif, in which two to seven alpha-helices are coiled together like the strands of a rope. Many coiled coils mediate oligomerization or protein–protein interaction, and the motif is important to the structure and function of several classes of fibrous structural proteins, motor proteins, transcription factors and membrane fusion proteins (Newman & Keating, 2003; Fong *et al.,* 2004).

Secondary structure prediction generally aims at correlating the frequencies of occurrence of short amino acid stretches with a particular secondary structure. The data set for this statistics is derived from known protein structures with the secondary structures assigned to the corresponding primary sequence. Characterization and identification of secondary structure of protein is often used as a constraint to tertiary structure prediction or as part of fold recognition methods (Russell *et al*., 1996). There are numerous secondary structure prediction methods such as PHD (Rost, 1996), PSIPRED (Jones, 1999), JPRED2 (Cuff & Barton, 2000, http://www.compbio.dundee.ac.uk/~www-jpred/), ZPRED (Zvelebil *et al.,* 1987, http://kestrel.ludwig.ucl.ac.uk/zpred.html), DSC (King & Sternberg, 1996, http://www.bmm.icnet.uk/dsc/dsc_form_align.html) and PREDATOR (Frishman & Argos, 1997,

http://www.embl-heidelberg.de/cgi/predator_serv.pl). The PHD and PSIPRED methods are widely used for the secondary structure prediction.

PHD: PHD is a suite of programs to predict 2-D structure (secondary structure, solvent accessibility) from multiple sequence alignment. The method scans the query sequence against SWISS-PROT database using BLASTP to identify similar sequences. A multiple sequence alignment is generated by a weighted dynamic programming method. Conserved motifs are retrieved from the PROSITE database. The evolutionary information from multiple alignments is used as input for profile-based neural network predictions (Rost, 1996). The average accuracy of PHD method is greater than 72% (http://www.predict protein.org/main.php).

PSIPRED: PSIPRED is a simple and accurate secondary structure prediction method, incorporating two feed-forward neural networks which perform an analysis based on position specific scoring matrices generated by PSI-BLAST (Jones, 1999). PSIPRED has maintained its position as one of the leading secondary structure prediction methods and found to be accurate with an average (Q3) score of 78% according to an independent continuous evaluation (Rost & Eyrich, 2001). PSIPRED was found to be reliable with ease of use and servicing over 15000 requests each month (Bryson *et al.,* 2005). The PSIPRED protein structure prediction server for sequence based searches is available at (http://bioinf.cs.ucl.ac.uk/psipred/psiform.html).

1.4 Analysis of hypothetical sequences: Often database searches using available sequence analysis tools do not yield results with respect to structural and functional information of genes or proteins. Such genes that have unknown function are called as orphan genes and code for proteins annotated as "Hypothetical proteins". However, more than 50% of proteins in the proteome zone remain unannotated and unidentified for function. Hence, there is a need

to begin constructing and analyzing protein families clustered as "Hypothetical proteins" with an aim to elucidate function and protein subunit interactions (Suravajhala, 2007). Although several databases explore protein functions through data-mining, there is a requirement to list all hypothetical proteins. There are reports that address the problem of orphan genes (Blayo *et al.,* 2003). An orphan gene is a gene that has no detectable homolog in other organisms with limited phylogenetic distribution. However, there is no adequate information to necessitate function of genes that cannot be based on homology alone, except connected to other known gene family. The identification of repeats and domains in proteins is one such approach which can better explain the functions for unannotated proteins or hypothetical proteins in the form of novel repeats and novel domains.

1.4.1 Repeats: A 'repeat' corresponds to a region of the protein sequence that occurs more than once in tandem, along the protein primary sequence. For example, the YVTN repeats in cell-surface proteins of several organisms. The tandem repeats can fold interdependently and form compact regular folding structures such as linear rods (eg. in spectrin) or superhelices (eg. HEAT repeats) or closed structures (β-propellers or β-trefoils).

Repeats are thought to arise due to gene duplication and recombination events. Unlike domains, repeats always exist in multiple copies (Andrade *et al.,* 2001; 2002). The repeat copy number and length may vary in different proteins indicating frequent loss or gain of these repeats during evolution. Repeats are often present in integer numbers and occasionally in non-integer numbers. When present as non-integers, the first half of a repeat is present at the C-terminus while the second half is present at the N-terminus. This mode of circular permutation in repeats was proposed for the SLH domain in eubacterial proteins (Lupas, 1996).

Information about the already identified repeats and domains is represented in the databases such as SMART, INTERPRO and PFAM. Some of the known repeats are as follows:

1. LRR Repeats: Leucine-rich repeats are present in diverse organisms that range from bacteria to human. They are present in over two hundred different proteins. They include hormone receptors, tyrosine kinase receptors, cell-adhesion molecules, bacterial virulence factors, enzymes and extracellular matrix binding glycoproteins (Matsushima *et al.,* 2000). The LRRs are usually present in tandem. The most common length of the LRR motif is 24 residues, but the lengths range from 20 to 30. All LRR motifs are divided into a highly conserved part and a variable part (Kajava *et al.,* 1995; Ohyanagi & Matsushima, 1997). The highly conserved part consists of a 11-residue stretch, LxxLxLxxNxL, or a 12-residue stretch, LxxLxLxxCxxL, where x is any amino acid residue (Ohyanagi & Matsushima, 1997). Many LRR proteins are involved in protein-ligand interactions; these include plant immune response and the mammalian innate immune response (Matsushima, 2005).

2. WD-40 Repeats: WD-40 repeats are minimally conserved domains of approximately 40–60 amino acids that are initiated by a glycine histidine (GH) dipeptide 11 to 24 residues from the N-terminus end with a tryptophan-aspartic acid (WD) dipeptide at the C-terminus. The repeating unit, first recognized in the β subunit of the GTP-binding protein transducin, has been referred as the transducin repeat, the GH-WD repeat, or the WD-40 repeat (Smith *et al.,* 1999; Neer *et al.,* 1994). Most WD-40 repeat proteins contain a cluster of at least 7 or more copies of the WD-40 repeats, with repeat numbers varying between 4 to 16. The WD-40 repeats are involved in signal transduction, transcriptional regulation and apoptosis. They are also associated with several human diseases. WD-40 repeat proteins are shown to form seven-bladed β propeller structure (Pickles *et al.,* 2000).

3. TPR Repeats: The tetratricopeptide repeat (TPR), a 34 amino acid motif, was first identified in yeast cell-division proteins (Sikorski *et al.,* 1990) and since been found in a variety of proteins associated with diverse biological functions. TPR repeats are commonly found in tandem arrays, typically with 3 to 16 direct repeats (D'Andrea & Regan, 2003). These arrays function as molecular scaffolds and frequently mediate protein–protein interactions (Main *et al.,* 2005). TPRs have been identified in >300 proteins, whose functions range from cell-cycle control to transcriptional regulation, protein transport, protein folding and neurogenesis (Blatch & Lassle, 1999; D'Andrea & Regan, 2003).

1.4.2 Domains: Domains are structural, functional and evolutionary units of the proteins (Murzin *et al.,* 1995; Holm & Sander, 1996; Orengo *et al.,* 1997). Domains have variety of definitions in different contexts. In crystallographer's definition, a domain is often viewed as a compact and spatially distinct folding unit. In biochemistry, domains are frequently described as protein regions with assigned experimental functions. In sequence comparison, domains are viewed from an evolutionary perspective and described as sequence regions with significant homology that are often present in different molecular contexts. However, these three views are compatible when sequence similar homologs adopt similar folds and exhibit comparable functions such as the domains in the signal transduction proteins SH2 (Src Homology 2), SH3 (Src Homology 3) and PH (Pleckstrin Homology). In the present context, a 'domain' refers to a region of the protein sequence that is present in a variety of other proteins and shares high sequence similarity and does not contain internal sequence repeats. Protein domains may exist either in multiple copies or a single copy per protein.

Domains can be readily observed in known 3-D structures, but because of the relative paucity of available structural data, the majority of protein domain families have been identified initially by sequence analysis. Many

domains are 'genetically mobile' and can be found to be associated with different domain combinations in a variety of proteins. Analysis of annotated domains provides clues in understanding the evolution of the domain classes. Novel domain identification in protein sequences helps in the classification of proteins into families by predicting the function and structures of a new protein or a poorly characterized protein and this can be achieved by sequence comparisons.

1. SH2 Domain: SH2 domains are protein modules (of ~100 amino acids) found in many proteins involved in tyrosine kinase signaling cascades. The structures of a large number of SH2 domains have been determined (Kuriyan & Cowburn, 1997). These studies have revealed a common fold consisting of a central β sheet flanked by 2 α helices. Their function is to bind tyrosine-phosphorylated sequences in specific protein targets. Binding of an SH2 domain to its cognate tyrosine-phosphorylated target links receptor activation to downstream signaling, both to the nucleus to regulate gene expression and throughout the cytoplasm of the cell (Waksman *et al.,* 2004).

2. SH3 Domain: SH3 region is a small protein domain (of ~56 amino acids) present in a very large group of proteins, including cytoskeletal elements and signaling proteins. It is believed that SH3 domains serve as modules that mediate protein-protein associations and along with Src homology 2 (SH2) domains regulate cytoplasmic signaling (Ren *et al.,* 1993). The SH3 domains were found to mediate protein-protein interactions by a proline-rich consensus sequence motif (Li, 2005).

3. PH Domain: PH domains comprise one of the largest domain families. They have been thoroughly investigated as modules that target membranes through recognition of phosphoinositide head groups (DiNitto *et al.,* 2006). Sequence profiles used to recognize PH domains primarily reflect their structural characteristics that can adopt (in ~100 amino acids) a 7 stranded β sheet structure with a C-terminal α helix (Lemmon & Ferguson, 2000).

1.4.3 Programs used for the Identification of Novel Repeats and Domains in Protein Sequences:

Several web-based methods are available for the *ab initio* identification of sequence repeats in proteins. The popular programs that identify internal repeats in proteins are REP Program (Andrade *et al.,* 2000), RADAR (Heger & Holm, 2000), REPRO (Heringa & Argos, 1993), PROSPERO (Mott, 2000) and TRUST (Szklarczyk & Heringa, 2004).

RADAR (Rapid **A**utomatic **D**etection and **A**lignment of **R**epeats**):** RADAR (Heger & Holm, 2000) uses an automatic algorithm for segmenting a query sequence into repeats, it identifies short composition biased as well as gapped approximate repeats and complex repeat architectures involving many different types of repeats in a query sequence (www.ebi.ac.uk/Radar).

REP (REPeat finding method): REP program (Andrade *et al.,* 2000) uses an iterative algorithm based on score distributions from profile analysis. This procedure allows the identification of homologs with alignment scores lower than the highest optimal alignment score for non-homologous sequences (http://www.embl- heidelberg.de/andrade/papers/rep/).

REPRO: REPRO program recognizes distant repeats in a single query sequence. The technique relies on a variation of Smith-Waterman local alignment strategy to find non-overlapping top-scoring local alignments, followed by a graph-based iterative clustering procedure to delineate the repeat set(s) based on consistency of the pair-wise top-alignments (Heringa & Argos, 1993). The program is available at (http://www.ibi.vu.nl/programs/reprowww/).

PROSPERO: The PROSPERO program (Mott, 2000) is ideal for large scale self comparison of protein sequences. It uses a formula that accurately assesses the significance of protein repeat similarities, allowing for existence of gaps and

also takes into account sequence length and composition. The program is available at (http://www.well.ox.ac.uk/ariadne/prospero.shtml).

TRUST (Tracking **R**epeats **U**sing **S**ignificance and **T**ransitivity**):** TRUST program (Szklarczyk & Heringa, 2004) exploits the concept of transitivity of alignments as well as a statistical scheme optimized for the evaluation of repeat significance. Starting from significant local sub-optimal alignments, the application of transitivity allows to: 1) identify distant repeat homologs for which no alignments were found; 2) gain confidence about consistently well aligned regions; and 3) recognize and reduce the contribution of non-homologous repeats. This assessment step will enable to derive a virtually noise free profile representing a generalized repeat with high fidelity. It has been demonstrated by the authors that TRUST is a useful and reliable tool for mining tandem and non-tandem repeats in protein sequence databases, to predict multiple repeat types with varying intervening segments within a single sequence. Once statistically significant repeats are detected, construction of a multiple sequence alignment provides insight into the extent of sequence homology among members of the new protein family and identification of the conserved sequence motifs. The TRUST server together with the source code is available at (http://ibivu.cs.vu.nl/programs/trustwww).

Detailed description for the repeat identification method is discussed in detail in Chapter 3.

1.5 Fold recognition methods: When sequence comparison methods are no longer sensitive enough to recognize structural homologs for a sequence, fold recognition methods are helpful in assigning the fold adopted by the sequence thereby detecting distantly related proteins. Protein folding is the physical process by which a polypeptide folds into its characteristic and functional 3-D structure (Bruce *et al.,* 2002). Some methods are based exclusively on sequence

information and other methods are based on multiple sequence alignment and structural information. Various methods used for fold prediction are FUGUE (Shi *et al*., 2001), INUB (Fischer *et al*., 2000, http://inub.bioinformatics.buffalo.edu/form.html), 3D-PSSM (3-D position-specific scoring matrix) (Kelley *et al*., 2000, http://www.sbg.bio.ic.ac.uk/~3dpssm/), FFAS **(**Rychlewski *et al*., 2000, http://bioinformatics.ljcrf.edu/FFAS/), GenTHREADER (Jones, 1999, http:// bioinf.cs.ucl.ac.uk/psipred/psiform.html), SPARKS 2.0 (Zhou & Zhou, 2004, http://sparks.informatics.iupui.edu), SP^3 (Zhou & Zhou, 2005, http://sparks. informatics.iupui.edu) and ROBETTA (Kim *et al.,* 2004, http://robetta. bakerlab.org). In our work, FUGUE was used for the fold prediction and is described below.

FUGUE: FUGUE is a program for recognizing distant homologs by sequence-structure comparison (Shi *et al*., 2001). It utilizes environment-specific substitution tables and structure-dependent gap penalties, where scores for amino acid matching and insertions/deletions are evaluated depending on the local environment of each amino acid residue in a known structure. Given a query sequence (or a sequence alignment), FUGUE scans a database of structural profiles, calculates the sequence-structure compatibility scores and produces a list of potential homologs and alignments. The prediction is evaluated on the basis of z score, which has to be ≥ 6.0 for a confident prediction of the fold. The FUGUE program is available at the website (http://www-cryst.bioc.cam.ac.uk/~fugue/prfsearch.html).

1.6 3-D Structure Modeling: The structures of proteins are being solved in increasing numbers, particularly as a result of structural genomics projects. Therefore, the number of protein structures that can be modeled are rising concomitantly (Baker & Sali, 2001). This structural information provides a basis for understanding protein function and for the design of modified proteins and ligands, including drugs (Harrison, 2004). Understanding the molecular

function of proteins is greatly enhanced by insights gained from their 3-D structures. Since experimental structures are only available for a small fraction of proteins, computational methods for protein structure modeling play an increasingly important role. Homology modeling is one such comparative structure prediction method that is widely used to build models of proteins with unknown structures based on the known structures of related proteins. Comparative protein structure modeling is currently the most accurate method, yielding models suitable for a wide spectrum of applications, such as structure-guided drug development or virtual screening (Kopp & Schwede, 2004).

1.6.1 Homology Modeling: Homology modeling, also known as comparative modeling, is a method for constructing an atomic-resolution model of a protein from its amino acid sequence (the "query sequence" or "target"). Homology modeling technique is based on the identification of one or more known protein structures (known as "templates" or "parent structures") likely to resemble the structure of the query sequence and on the production of an alignment that maps residues in the query sequence to residues in the template sequence. The sequence alignment and template structure are then used to produce a structural model of the target. It is generally accepted that proteins with high sequence similarity also possess structural similarity (Marti-Renom *et al.,* 2000). For proteins that share greater than 50% sequence identity, the root mean square deviation (RMSD) of the alpha-carbon co-ordinates is observed to be less than 1Å.

The homology modeling procedure is carried out in four sequential steps: template selection, target-template alignment, model construction and model assessment. In order to identify the template structures, target sequence is searched against the Protein Data Bank (PDB), using programs such as FASTA and BLAST. The best template structure will be the one with the highest sequence similarity to the target and will serve as the template.

Homology modeling is a powerful technique that greatly increases the value of experimental structure determination by using the structural information of one protein to predict the structures of homologous proteins (Bhattacharya *et al.,* 2008). Several methods are in use for homology modeling, here we present a brief discussion of MODELLER, a module in homology of INSIGHTII (version 2000, Accelrys, Sandiego, CA).

MODELLER: MODELLER is a well known computer program for comparative protein structure modeling. It takes the sequence alignment between the target sequence and template structure as input and produces a comparative model. MODELLER implements comparative protein structure modeling by satisfaction of spatial restraints (Sali & Blundell, 1993). The spatial restraints include homology-derived restraints on the distances and dihedral angles in the target sequence extracted from its alignment with the template structures (Sali & Blundell, 1993); stereochemical restraints such as bond length and bond angle preferences obtained from the CHARMM22 molecular mechanics forcefield (Mac Kerell *et al*., 1998); statistical preferences for dihedral angles and non-bonded interatomic distances obtained from a representative set of known protein structures (Sali & Overington, 1994). MODELLER provides an option to curate the restraints manually, such as those from NMR spectroscopy, rules of secondary structure packing, cross-linking experiments, fluorescence spectroscopy, image reconstruction from electron microscopy and site-directed mutagenesis. The spatial restraints, expressed as probability density functions, are combined into an objective function that is optimized by a combination of conjugate gradients and molecular dynamics with simulated annealing (Eswar *et al*., 2003).

1.6.2 3-D Structure Validation: The explosive increase in the number of published 3-D structures of macromolecules determined by X-ray analysis

places a responsibility on experimentalists, referees and curators of databases to ensure correspondence between the structure parameters and data (Dodson *et al.,* 1998). Protein structures that are derived from either experimental data or computational predictions are mere model structures. These models aim to give reasonable explanation for the input data such as diffraction pattern or NMR restraints (Laskowski, 2003). The quality, quantity and care with which the data was collected, reflects the accuracy of the protein model built with this data. It is well known that in X-ray crystallography, NMR and especially in protein structure prediction, errors can be introduced at various stages of the model building process, which is well documented in literature (Branden & Jones, 1990; Hooft *et al*., 1996). Also in theoretical protein modeling, misalignment of amino acids with respect to the true position in the fold can seriously mislead the functional interpretation. To surmount these problems various methods have been developed for protein structure validation. These methods evaluate the stereo chemical quality and sequence structure correlation of protein models.

A concise description of PROCHECK and profiles-3D validation methods is presented below.

PROCHECK: PROCHECK is a suite of programs that offers a detailed analysis on the stereochemistry of a protein structure (Laskowski *et al*., 1993). The program verifies a variety of geometry-based criteria such as Ramachandran plot (Ramachandran, *et al.,* 1963), main chain, side chain, bond lengths and angles, planarity of rings and end groups, torsion angles, chirality, close non-bonded interactions, main chain H-bonds, disulfide bond geometry and residue by residue analysis. Accordingly, it generates a number of postscript plots analyzing its overall and residue-by-residue geometry.

Profiles-3D: Profiles-3D examines the validity of a preliminary structure or model derived from experimental data or modeling studies. It measures the compatibility between the protein sequences and known protein structures. Profiles-3D evaluates the 3-D structure by comparing its structural environments with the preferred environments of the amino acids in the known sequences. Environment is defined by the following criteria (i) the area of the residue that is buried; (ii) the fraction of side-chain area that is covered by polar atoms (oxygen and nitrogen); (iii) the local secondary structure. If a residue lies in an unusual chemical environment, it will receive a bad score and vice versa. Given a 3-D structure, it identifies which amino acid sequences are compatible with that structure (Luethy *et al*., 1992).

1.6.3 3-D Structural Database:

PDB (Protein Data Bank): The tertiary structure of a protein or any other macromolecule is its 3-D structure, as defined by the atomic coordinates determined by the protein's primary sequence. All the known 3-D structural data of biological macromolecules are deposited at PDB (Dutta *et al*., 2008) which is an important database source that provides access to the 3-D co-ordinates and related information of the biological macromolecules that help in understanding the folding pattern, ligand binding etc. of these molecules. This structural information is exploited in protein classification as well as drug design studies. The fast growing PDB contains the 3-D description of more than 51366 proteins and nucleic-acid structures. The database is made available to researchers worldwide via the website (www.rcsb.org/pdb).

1.7 Docking: The binding of small molecule ligands to large protein targets is central to numerous biological processes. The accurate prediction of the binding modes between the ligand and protein (the docking problem) is of fundamental importance in modern structure-based drug design. Molecular docking is

defined as an optimization process, which would describe the "best-fit" orientation of a ligand that binds to a particular protein of interest.

Docking is frequently used to predict the binding orientation of small molecule drug candidates to their protein targets in order to be able to in turn predict the affinity and activity of the small molecule. Hence docking plays an important role in the rational drug design studies (Kitchen *et al.,* 2004). A good docking method places the ligand appropriately in the active site and then estimates the forces involved in the receptor-ligand recognition (electrostatic, van der Waals and hydrogen bonding).

Docking comprises of two components 1. Configurational and conformational degrees of freedom, 2. Scoring function. The search algorithm searches the potential energy landscape adequately to find the global energy minimum. In rigid docking, the search algorithm explores different positions for the ligand in the receptor active site using the translational and rotational degrees of freedom. Flexible ligand docking adds exploration of torsional degrees of freedom of the ligand. These algorithms are complemented by scoring functions that are designed to predict the biological activity through the evaluation of interactions between compounds and potential targets. The scoring function has to be realistic enough to assign the most favorable scores to the experimentally determined complex. Usually, the scoring function assesses both the steric as well as the chemical complementarities between the ligand and the receptor.

The process of evaluating the particular conformation of molecule when bound to protein uses a number of descriptive features such as, number of intermolecular interactions including hydrogen bonds, hydrophobic contacts and van der Waals energy. Scoring function used in docking is a mathematical function whose values are proportional to the binding affinities of the lead molecules. A good scoring function should be able to give reliable estimates of binding affinities of structurally diverse lead molecules for different protein

targets while considering the thermodynamic aspects of binding (Ajay & Murko, 1995). The success of a docking program depends on both the search algorithm and the scoring function.

Docking is most commonly used in the field of drug design for two purposes, 1. hit identification: docking combined with a scoring function can be used to quickly screen large databases of potential drugs *in silico* to identify molecules that are likely to bind to the protein target of interest and 2. lead optimization: to predict the correct location and relative orientation of a ligand binding to a protein. This information may in turn be used to design more potent and selective analogs.

The problems and methods introduced in this chapter have been instrumental in the advance of our understanding of protein organization structure and function. The computational tools aimed at analyzing the protein data are useful in identifying novel repeats and domains in proteins.

1.8 References

Ajay & Murko, M. A. (1995). Computational methods to predict binding free energy in ligand-receptor complexes. *J. Med. Chem.* **38,** 4953-4967.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. (1990). Basic local alignment search tool. *J. Mol. Biol.* **215,** 403-410.

Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucl. Acids Res.* **25,** 3389-3402.

Andrade, M. A., Ponting, C. P., Gibson, T. J. & Bork, P. (2000). Homology-based method for identification of protein repeats using statistical significance estimates. *J. Mol. Biol.* **298,** 521-537.

Andrade, M. A., Perez-Iratxeta, C. & Ponting, C. P. (2001). Protein repeats: structures, functions, and evolution. *J. Struct. Biol.* **134**, 117-131.

Andrade, M. A., Ciccarelli, F. D., Perez-Iratxeta, C. & Bork, P. (2002). NEAT: a domain duplicated in genes near the components of a putative Fe^{3+} siderophore transporter from Gram- positive pathogenic bacteria. *Genome Biol.* **3**, 0047.1- 0047.5.

Baeza-Yates, R. & Gonnet, G. (1992). A New Approach to Text Searching. *Commun. Assoc. Comp. Mach*. **35,** 74–82.

Bairoch, A. & Apweiler, R. (2000). The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res.* **28,** 45-48.

Baker, D. & Sali, A. (2001). Protein structure prediction and structural genomics. *Science,* **294,** 93–96.

Bateman, A., Coin, L., Durbin, R., Finn, R. D., Hollich, V., Griffiths-Jones, S., Khanna, A. *et al*. (2004). The Pfam protein families database. *Nucleic Acids Res.* **32,** D138-D141.

Bhattacharya, A., Wunderlich, Z., Monleon, D., Tejero, R. & Montelione, G. T. (2008). Assessing model accuracy using the homology modeling automatically software. *Proteins,* **70,** 105-118.

Blatch, G. L. & Lassle, M. (1999). The tetratricopeptide repeat: a structural motif mediating protein–protein interactions. *BioEssays,* **21,** 932–939.

Blayo, P., Rouz-e, P. & Sagot, M. F. (2003). Orphan gene finding-an exon assembly approach. *Theoretical Computer Science,* **290,** 1407–1431.

Branden, C. I. & Jones, T. A. (1990). Between objectivity and subjectivity. *Nature*, **343**, 687-689.

Bruce, A., Johnson, A., Lewis, J., Raff, M., Roberts, K. & Walters, P. (2002). "The Shape and Structure of Proteins", *Molecular Biology of the Cell, Fourth Edition*. New York and London: Garland Science. ISBN 0-8153-3218-1.

Bryson, K., McGuffin, L. J., Marsden, R. L., Ward, J. J, Sodhi, J. S & Jones, D. T. (2005). Protein structure prediction servers at University College London. *Nucleic Acids Research*, **33,** W36–W38.

Cuff, J. A. & Barton, G. J. (2000). Application of multiple sequence alignment profiles to improve protein secondary structure prediction. *Proteins,* **40,** 502-511.

D'Andrea, L. D. & Regan, L. (2003). TPR proteins: the versatile helix. *Trends Biochem. Sci.* **28,** 655–662.

Dayhoff, M. O. (1978). Survey of new data and computer methods of analysis. In *Atlas of protein sequence and structure*, pp.1-8, *Nat. Biomed. Res. Found*., Washington D.C.

DiNitto, J. P. & Lambright, D. G. (2006). Membrane and juxtamembrane targeting by PH and PTB domains. *Biochimica et Biophysica Acta,* **1761,** 850–867.

Dodson, E. J., Davies, G. J., Lamzin, V. S., Murshudov, G. N. & Wilson. K. S. (1998). Validation tools: can they indicate the information content of macromolecular crystal structures? *Structure,* **6,** 685-690.

Dutta, S., Burkhardt, K., Swaminathan, G. J., Kosada, T., Henrick, K., Nakamura, H. & Berman, H. M. (2008). Data deposition and annotation at the worldwide protein data bank. *Methods Mol Biol.* **426,** 81-101.

Eswar, N., John, B., Mirkovic, N., Fiser, A., Ilyin, V. A., Pieper, U., Stuart, A. C. *et al*. (2003). Tools for comparative protein structure modeling and analysis. *Nucl. Acids Res.* **31**, 3375-3380.

Feng, D. F. & Doolittle, R. F. (1987). Progressive sequence alignment as a prerequisite to correct phylogenetic trees. *J. Mol. Evol.* **25**, 351-360.

Finn, R. D., Tate, J., Mistry, J., Coggill, P. C., Sammut, S. J., Hotz, H. R., Ceric, G. *et al.* (2008). The Pfam protein families database. *Nucleic Acids Res.* **36,** Database issue, D281-289.

Fischer, D. (2000). Hybrid Fold Recognition: Combining Sequence Derived Properties with Evolutionary Information. *Pacific Symp. Biocomput.* 119-130.

Fong, J. H., Keating, A. E. & Singh M. (2004). Predicting specificity in bZIP coiled–coil protein interactions. *Genome Bio.* **5,** R11.

Fraser, A. G. & Marcotte, E. M. (2004). A probabilistic view of gene function. *Nat. Genet.* **36,** 559-564.

Frishman, D. & Argos, P. (1997). Seventy-five percent accuracy in protein secondary structure prediction. *Proteins*, **27**, 329-335.

Harrison, S. C. (2004). Whither structural biology? *Nat. Struct. Mol. Biol.* **11**, 12–15.

Heger, A. & Holm, L. (2000). Rapid automatic detection and alignment of repeats in protein sequences. *Proteins,* **41,** 224–237.

Henikoff, S. & Henikoff, J. G. (1992). Amino acid substitution matrices from protein blocks. *Proc. Natl. Acad. Sci.* **89,** 10915-10919.

Heringa, J. & Argos, P. (1993). A method to recognize distant repeats in protein sequences. *Proteins*, **17,** 391–341.

Hocquette, J. F. (2005). Where are we in genomics? J*ournal of Physiology and Pharmacology*, **56**, 37-70.

Holm, L. & Sander, C. (1996). Mapping the protein universe. *Science*, **273**, 595–603.

Hooft, R. W., Vriend, G., Sander, C. & Abola, E. E. (1996). Errors in protein structures. *Nature*, **381**, 272.

Hulo, N., Bairoch, A., Bulliard, V., Cerutti, L., Castro, D. E., Petra, S., Genevaux, L. *et al.* (2006). The PROSITE database. *Nucleic Acids Research,* **34,** Database issue, D227–D230.

Innis, C. A., Shi, J. & Blundell, T. L. (2000). Evolutionary Trace Server (TraceSuite II). *Protein Eng.* **13,** 839-847.

Jones, D. T. (1999). Protein secondary structure prediction based on position-specific scoring matrices. *J. Mol. Biol.* **292,** 195-202.

Jones, D. T. (1999). GenTHREADER: an efficient and reliable protein fold recognition method for genomic sequences. *J. Mol. Biol.* **287**, 797-815.

Kajava, A. V., Vassart, G. & Wodak, S. J. (1995). Modeling of the three-dimensional structure of proteins with the typical leucine-rich repeats. *Structure*, **3,** 867–877.

Kelley, L. A., Mac Callum, R. M. & Sternberg, M. J. E. (2000). Enhanced genome annotation using structural profiles in the program 3D-PSSM. *J. Mol. Biol.* **299,** 499-520.

Kim, D. E., Chivian, D. & Baker, D. (2004). Protein structure prediction and analysis using the Robetta server. *Nucleic Acids Res.* **32**, W526-W531.

King, R. D. & Sternberg, M. J. (1996). Identification and application of the concepts important for accurate and reliable protein secondary structure prediction. *Protein Sci.* **11**, 2298-2310.

Kitchen, D. B., Decornez, H., Furr, J. R. & Bajorath, J. (2004). Docking and scoring in virtual screening for drug discovery: methods and applications. *Nat. Rev. Drug Discov.* **3,** 935-949.

Kopp, J. & Schwede, T. (2004). Automated protein structure homology modeling: a progress report. *Pharmacogenomics,* **5,** 405-416.

Kuriyan, J. & Cowburn, D. (1997). Modular peptide recognition domains in eukaryotic signaling. *Annu. Rev. Biophys. Biomol. Struct.* **26,** 259-288.

Laskowski, R. A., Mac Arthur, M. W., Moss, D. S. & Thornton, J. M. (1993). PROCHECK: a program to check the stereo chemical quality of protein structures. *J. Appl. Crystallog.* **26,** 283-291.

Laskowski, R. A. (2003). Structural quality assurance. *Methods Biochem. Anal.* **44,** 273-303.

Letunic, I., Copley, R. R., Pils, B., Pinkert, S., Schultz, J. & Bork, P. (2006). SMART 5: domains in the context of genomes and networks. *Nucleic Acids Res.* **34**, Database issue, D257-260.

Lemmon, M. A. & Ferguson, K. M. (2000). Signal-dependent membrane targeting by pleckstrin homology (PH) domains. *Biochem. J*. **350,** 1–18.

Li, S. S. (2005). Specificity and versatility of SH3 and other proline-recognition domains: structural basis and implications for cellular signal transduction. *Biochem. J*. **390**, 641–653.

Lichtarge, O., Bourne, H. R. & Cohen, F. E. (1996). An evolutionary trace method defines binding surfaces common to protein families. *J. Mol. Biol.* **257,** 342–358.

Lin, J. & Qian, J. (2007). Systems biology approach to integrative comparative genomics. *Expert Rev. Proteomics,* **4,** 107-119.

Luethy, R., Bowie, J. U. & Eisenberg, D. (1992). Assessment of protein models with three-dimensional profiles. *Nature*, **356**, 83-85.

Lupas, A. (1996). A circular permutation event in the evolution of the SLH domain? *Mol. Microbiol.* **20**, 897-898.

Mac Kerell Jr., A. D. J., Bashford, D., Bellott, R. L., Dunbrack Jr., R. L., Evanseck, J. D., Field, M. J., Fischer, S. *et al*. (1998). All-Atom empirical potential for molecular modeling and dynamics studies of proteins. *J. Phys. Chem. B.* **102**, 3586–3616.

Main, E. R., Stott, K., Jackson, S. E. & Regan, L. (2005). Local and long-range stability in tandemly arrayed tetratricopeptide repeats. *Proc. Natl. Acad. Sci.* USA, **102,** 5721–5726.

Marti-Renom, M. A., Stuart, A. C., Fiser, A., Sanchez, R., Melo, F. & Sali, A. (2000). Comparative protein structure modeling of genes and genomes. *Annu. Rev. Biophys. Biomol. Struct.* **29,** 291-325.

Matsushima, N., Kamiya, M., Suzuki, N. & Tanaka, T. (2000). Super-Motifs of Leucine-Rich Repeats (LRRs) Proteins. *Genome Informatics,* **11,** 343-345.

Matsushima, N., Enkhbayar, P., Kamiya, M., Osaki, M. & Kretsinger, R. H. (2005). Leucine-rich repeats (LRRs): structure, function, evolution and interaction with ligands. *Drug Design Reviews*, **4,** 305-322.

Mott, R. (2000). Accurate formula for P-values of gapped local sequence and profile alignments. *J. Mol. Biol.* **300,** 649-659.

Mulder, N. J., Apweiler, R., Attwood, T. K., Bairoch, A., Bateman, A., Binns, D., Bradley, P. *et al*. (2005). InterPro, progress and status in 2005. *Nucleic Acids Res.* **33,** D201-D205.

Mulder, N. J., Apweiler, R., Attwood, T. K., Bairoch, A., Bateman, A., Binns, D., Bork, P. *et al*. (2007). New developments in the InterPro database. *Nucleic Acids Res.* **35,** Database issue, D224-D232.

Murzin, A. G., Brenner, S. E., Hubbard, T. & Chothia, C. (1995). SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* **247**, 536-540.

Neer, E. J., Schmidt, C. J., Nambudripad, R. & Smith, T. F. (1994). The ancient regulatory-protein family of WD-repeat proteins. *Nature,* **371,** 297–300.

Newman, J. R. & Keating, A. E. (2003). Comprehensive identification of human bZIP interactions with coiled–coil arrays. *Science,* **300,** 2097–2101.

Ohyanagi, T. & Matsushima, N. (1997). Classification of tandem leucine-rich repeats within a great variety of proteins. *FASEB J.* **11,** A949.

Orengo, C. A., Michie, A. D., Jones, S., Jones, D. T., Swindells, M. B. & Thornton, J. M. (1997). CATH-a hierarchic classification of protein domain structures. *Structure*, **5,** 1093-1108.

Pearson, W. R. & Lipman, D. J. (1988). Improved Tools for Biological Sequence Analysis. *Proc. Natl. Acad. Sci. USA*. **85,** 2444-2448.

Pickles, L. M., Roe, S. M., Hemingway, E. J., Stifani, S. & Pearl, L. H. (2000). Crystal structure of the C-terminal WD40 repeat domain of the human Groucho/TLE1 transcriptional corepressor. *Structure,* **6**, 751-761.

Ramachandran, G. N., Ramakrishnan, C. & Sasisekharan, V. (1963). Stereochemistry of polypeptide chain configurations. *J. Mol. Biol.* **7**, 95-99.

Ren, R., Mayer, B. J, Cicchetti, P. & Baltimore, D. (1993). Identification of a ten-amino acid proline-rich SH3 binding site. *Science*, **259,** 1157-1161.

Rost, B. (1996). PHD: predicting one-dimensional protein structure by profile based neural networks. *Methods Enzymol.* **266,** 525-539.

Rost, B. & Eyrich, V. A. (2001). EVA: large-scale analysis of secondary structure prediction. *Proteins*, **5,** 192–199.

Russell, R. B., Copley, R. R. & Barton, G. J. (1996). Protein fold recognition by mapping predicted secondary structures. *J. Mol. Biol.* **259**, 349-365.

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

Sali, A. & Blundell, T. L. (1993). Comparative protein modeling by satisfaction of spatial restraints. *J. Mol. Biol.* **234**, 779-815.

Sali, A. & Overington, J. P. (1994). Derivation of rules for comparative protein modeling from a database of protein structure alignments. *Protein Sci.* **3**, 1582–1596.

Shi, J., Tom, L. B. & Kenji, M. (2001). FUGUE: Sequence structure homology recognition using environment-specific substitution tables and structure-dependent gap penalties. *J. Mol. Biol.* **310,** 243-257.

Sikorski, R. S., Boguski, M. S., Goebl, M. & Hieter, P. (1990). A repeating amino acid motif in CDC23 defines a family of proteins and a new relationship among genes required for mitosis and RNA synthesis. *Cell,* **60,** 307–317.

Smith, T. F. & Waterman, M. S. (1981a). Identification of common molecular subsequences. *J. Mol. Biol.* **147**, 195-197.

Smith, T. F. & Waterman, M. S. (1981b). Comparison of biosequences. *Adv. Appl. Math.* **2**, 482-489.

Smith, T. F., Gaitatzes, C., Saxena, K. & Neer, E. J. (1999). The WD repeat: a common architecture for diverse functions. *Trends Biochem. Sci.* **24,** 181–185.

Suravajhala, P. (2007). Hypo, hype and 'hyp' human proteins. *Bioinformation,* **2,** 31-33.

Szklarczyk, R. & Heringa, J. (2004). TRUST: Tracking Repeats Using Significance and Transitivity. *Bioinformatics,* **00,** 1-7.

The UniProt Consortium. (2008). The Universal Protein Resource (UniProt). *Nucleic Acids Research,* **36,** Database issue, D190–D195.

Thompson, J. D., Higgins, D. G. & Gibson, T. J. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucl. Acids Res.* **22,** 4673-4680.

Waksman, G., Kumaran, S. & Lubman, O. (2004). SH2 domains: role, structure and implications for molecular medicine. e*xpert reviews in molecular medicine,* **6,** 1-21.

Wilbur, W. J. & Lipman, D. J. (1983). Rapid similarity searches of nucleic acid and protein data banks. *Proc. Natl. Acad. Sci. USA*. **80**, 726-730.

Wu, S. & Manber, U. (1992). Fast Text Searching Allowing Errors. *Commun. Assoc. Comp. Mach*. **35**, 83–91.

Zhang, Z., Schäffer, A. A., Miller, W., Madden, T. L., Lipman, D. J., Koonin, E. V. & Altschul, S. F. (1998). Protein sequence similarity searches using patterns as seeds. *Nucleic Acids Res*. **26,** 3986-3990.

Zhang, C. & Kim, S. H. (2003). Overview of structural genomics: from structure to function. *Curr. Opin. Chem. Biol.* **7,** 28–32.

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

Zhou, H. & Zhou, Y. (2005). Fold Recognition by Combining Sequence Profiles Derived From Evolution and From Depth-Dependent Structural Alignment of Fragments. *Proteins*, **58,** 321–328.

Zvelebil, M. J. J. M., Barton, G. J., Taylor, W. R. & Sternberg, M. J. E. (1987). Prediction of Protein Secondary Structure and Active Sites Using the Alignment of Homologous Sequences. *Journal of Molecular Biology,* **195,** 957-961.

CHAPTER 2

Analysis, 3-D structure modeling, Docking and Gene Cluster Identification of CMN mycolyl-transferases

2.1 Introduction

Tuberculosis assumes perilous trends in synergy with HIV infection, which has led the WHO to declare TB as global health emergency. *Mycobacterium tuberculosis*, the primary etiological agent of tuberculosis (TB) affects one-third of the world's population (Tabbara, 2007). *M. tuberculosis* is surrounded by a complex cell envelope which consist of three parts 1) a plasma membrane, 2) an asymmetric lipid bilayer - the inner layer consists of covalently linked three-component superpolymer of mycolic acids, D-arabino-D-galactan and peptidoglycan (mAGP) complex and the outer layer consists of non-covalently linked α, α' trehalosedimycolate (TDM, commonly referred as cord factor), α, α' trehalosemonomycolate (TMM) and other lipids and 3) an outer most layer, which is also called capsule, consists of mainly polysaccharides and proteins along with small amount of lipids (Daffé & Draper, 1998).

2.1.1 CMN group: The CMN group constitutes the organisms of the genera Corynebacterium, Mycobacterium and Nocardia, which are grouped together on the basis of factors that include complex cell wall components, presence and type of mycolic acids, adjuvant activity, presence of cord factor, sulfo-lipids, iron-chelating compounds, polyphosphate and serological cross-reactivity. The cell walls of the organisms that belong to the CMN group consists of interconnected peptidoglycan and polysaccharide-mycolate complex and are characterized by the presence of mycolic acid on their surface (Cocito & Delville, 1985).

The genome sequencing of *M. tuberculosis* (Cole *et al.,* 1998), *C. glutamicum,* (Kalinowski *et al.,* 2003), *C. efficiens* (Kawarabayasi, *et al.,* 2002), *C. diphtheria* (Tarraga *et al.,* 2003) and *N. farcinica* (Ishikawa *et al.,* 2004) is completed. The *M. tuberculosis* is the causative agent of tuberculosis, it

consists of 3,986 genes with 65.6% G+C content. The *C. glutamicum* is a soil bacterium and widely used by the industry in the production of amino acids. It consists of 3,002 genes with 53.8% G+C content. The *C. efficiens* is a non-pathogenic bacterium and consists of 3,069 genes and 63.4% G+C content. The *C. diphtheria* is the causative agent of diphtheria and consists of 2,320 genes with 53.48% G+C content. The genome of *N. farcinica*, the causative agent of nocardiosis, affecting the lung, central nervous system and cutaneous tissues of humans and animals consists of 5,674 genes with 70.8% G+C content.

2.1.2 Mycolic Acids: Mycolic acids are long chain α-alkyl, β-hydroxyl fatty acids that form a part of the unique cell envelope, responsible for the pathogenesis and survival of the organism inside the host. They play a crucial role in the biogenesis and organization of cell wall and also in numerous biological functions related to both the physiology and the virulence of mycobacteria (Brennan & Nikaido**,** 1995; Draper, 1998). Particularly, trehalose mycolates aid in virulence, where the structure of the mycolates has been found to be important for initial replication and persistence *in vivo* (Glickman *et al*., 2000).

The mycolic acids are named according to the individual genus from which they are isolated; i.e., corynemycolic acids from Corynebacterium comprising ~22-36 carbons, mycolic/eumycolic acids from Mycobacterium comprising ~60-90 carbons and nocardiomycolic acids from Nocardia comprising ~40-60 carbons (Collins *et al.,* 1982; Alashamaony *et al.,* 1976). The occurrence of mycolic acids is limited to the cell envelopes of corynebacteria, mycobacteria, nocardia, rhodococcus and related taxa that are collectively called CMN group (Minnikin *et al*., 1978).

2.1.3 Mycolyl-transferases: Mycolyl-transferases were first identified in *M. tuberculosis* and they are also termed as antigen 85 (Ag85) complex enzymes.

These correspond to three secreted proteins; Ag85A (GENE_ID: Rv3804), Ag85B (GENE_ID: Rv1886) and Ag85C (GENE_ID: Rv0129) (Wiker & Harboe, 1992). These proteins catalyze the transfer of the mycolic acid and comprise a signal peptide at the N-terminus followed by a carboxylesterase domain. It has been demonstrated that Ag85 enzymes catalyse the transfer of mycolyl residue from one molecule of α, α' – TMM (trehalose monomycolate) to another leading to the formation of α, α' – TDM (trehalose dimycolate) and hence these enzymes are termed mycolyl-transferases (Belisle, *et al.,* 1997). Also, in Corynebacterium and Nocardia, orthologous proteins synthesize TDCM (trehalose dicorynemycolate) and TDNM (trehalose dinocardio mycolate), respectively. Further, this family of enzymes are specific only to the CMN group of organisms because of their unique cell envelope. Mycolyl-transferases are also termed fibronectin-binding proteins, since they are involved in binding to fibronectin and aids in the entry of the organism into host cells (Abou-Zeid *et al.,* 1988; Ratliff *et al.,* 1988). Hence, it is important to understand the structure and function of the proteins responsible for the synthesis of cell wall components in CMN.

Mycobacterial DNA-binding protein1 (MDP1) which was designated as a histone like DNA binding protein1, plays an important role on mycolyl-transferase functions of the Ag85 complex through direct binding to both the Ag85 complex and the substrate, trehalose-6-monomycolate, in the cell wall (Katsube *et al.,* 2007).

The structures of Ag85A (PDB ID: 1SFR) (Ronning *et al.,* 2004), Ag85B (PDB IDs: 1F0N, 1F0P) (Anderson *et al.,* 2001) and Ag85C (PDB IDs: 1DQZ, 1DQY, 1VA5) (Ronning *et al.,* 2000) were determined for both native and substrate bound forms. The structure corresponds to a α/β hydrolase fold and the catalytic triad responsible for the mycolyl-transferase activity comprises the amino acid residues S126, E230 and H262 (numbering according to the

PDB ID: 1F0P). The structural comparison of the three mycolyl-transferases (PDB IDs: 1SFR, 1F0P, 1DQZ) revealed that the active sites are virtually identical, indicating that these share a common function (Ronning *et al.,* 2004). However, in contrast to the high level of similarity within the substrate-binding site and the active site, it was observed that the surface residues disparate from the active site are quite variable, indicating that all the three Ag85 enzymes in *M. tuberculosis* are needed to evade the host immune system.

In our earlier work (Adindla *et al.,* 2004a), we identified mycolyl-transferases in *C. glutamicum* and *C. efficiens* genomes and modeled their 3-D structures. We reported the relative binding of corynemycolyl-transferases towards trehalose. Our findings are in accordance with the experimental data (Brand *et al*., 2003; De Sousa *et al.,* 2003) that reported the gene deletion mutation studies and measured the concentration of TMCM / TDCM.

The genomes of *N. farcinica*, a representative species from Nocardia, and *C. diphtheria* were also subsequently sequenced. We now have complete data available in the public databases on all the mycolyl-transferases from species that belong to the CMN group, since the mycolyl-transferases are present to these organisms.

2.1.4 Gene cluster analysis: The availability of multiple, complete genomes of diverse life forms for comparative analysis provides a qualitatively new perspective on homologous relationships between genes. By comparing the sequences of all genes between genomes from different taxa and within each genome, it is possible to reconstruct the evolutionary history of each gene in its entirety (within the set of sequenced genomes). This, inturn, will allow a deeper understanding of the general trends in the evolution of genomic complexity and lineage-specific adaptations (Koonin, 2005). Gene histories are presented in the form of scenarios and comprise of several types of elementary events (Kunin & Ouzounis, 2003; Mirkin *et al.,* 2003; Snel *et al.,* 2002). The elementary events

of gene evolution can be classified roughly in the order of relative contribution to the evolutionary process as follows: (*i*) vertical descent (speciation) with modification; (*ii*) gene duplication, also followed by descent with modification; (*iii*) gene loss; (*iv*) horizontal gene transfer (HGT) and (*v*) fusion, fission and other rearrangements of genes. Vertical descent and duplication might be considered the primary events of genome evolution and have been well recognized in the pregenomic era. In contrast, the crucial evolutionary importance of gene loss, HGT and gene rearrangements were among the major, fundamental generalizations of the emerging evolutionary genomics (Doolittle 1998; 1999; 2000; Koonin & Galperin, 2002; Koonin, 2001; Lawrence & Hendrickson, 2003; Pennisi, 1998; 2001).

Genome-scale mapping of orthologs and paralogs is considered to be the first step when studying the evolution of proteins with shared ancestry, interactions and regulation (Wapinski *et al.,* 2007). A brief description of each of the important terminologies like homologs, orthologs, paralogs and operons is given below.

Homologs are the genes that share a common origin.

Orthologs are the genes in different species that are evolved from a common ancestral gene by speciation. The encoded proteins generally are 60-80% identical in sequence. Normally, orthologs retain the same function in the course of evolution (Ex: alpha hemoglobin in man and mouse).

Paralogs are the genes related by duplication within a genome. Paralogs may sometimes evolve more specific functions that are related to the original one (Ex: alpha and beta hemoglobin).

Operons are clusters of co-transcribed genes that often encode functionally linked proteins. These are the principal form of the gene organization and regulation in prokaryotes.

Several bioinformatics methods have recently been developed to analyze the genomic context of genes. These include analysis of a) gene fusions (Enright *et*

al., 1999; Marcotte *et al*., 1999), b) gene clusters (Overbeek *et al*., 1999), c) gene neighborhood (assembly of genes in putative operons) (Dandekar *et al*., 1998; Overbeek *et al*., 1999), and d) co-occurrence of genes across the genomes (Pellegrini *et al*., 1999) to predict functional associations for a given protein such as physical interaction partners or members of the same biological pathway. The genomic context methods provide a new and important development in genomics that explicitly takes advantage of the rapidly growing collection of sequenced genomes.

a) Gene fusions: Gene fusion occurs by gene recombination events, which results in one long composite protein in one of the orthologs, and two or more smaller split or component proteins in another ortholog. Gene fusion leads to the formation of multidomain proteins. It is a well-known process in the molecular evolution (Doolittle, 1999). Detection of gene fusions in one genome allows the prediction of physical interactions and functional associations between homologous genes that remain separate in another genome (Enright *et al*., 1999; Marcotte *et al*., 1999; Huynen *et al*., 2000).

b) Gene clusters: Gene clusters are defined as a set of close proximal genes that are functionally related (Overbeek *et al*., 1999). Gene clusters that are conserved among diverse bacterial genomes are known as operons. Functionally related genes often tend to cluster as co-transcribed and co-regulated operons. The physical proximity of genes in operons infers a physical interaction between the corresponding proteins or involvement in the same metabolic pathway (Lawrence, 1997; Overbeek *et al*., 1999). If a given metabolic pathway is important for the survival of an organism all the components of the pathway will be conserved in the organism and if the pathway becomes dispensable, all genes will tend to disappear (Marcotte *et al*., 1999). The occurrence of genes in the same neighborhood in multiple phylogenetically distant genomes is a strong indication of functional interactions between their proteins. This reflects a physical interaction between

the corresponding proteins or an involvement in the same metabolic pathway (Dandekar *et al*., 1998; Overbeek *et al*., 1999; Huynen *et al.,* 2000). In prokaryotes, the genes that are functionally related are located in close proximity in the genome and not necessarily in an operon.

To gain insights into the function and evolution of these proteins we have therefore carried out the sequence analysis, 3-D modeling of the structures of mycolyl-transferases in related genomes using the homology modeling method, docking of trehalose substrate into the binding site of all protein models, evolutionary trace analysis, genomic context analysis and comparison of the substrate binding sites. This analysis is relevant in situations when the structural information for proteins from only one organism are available and useful inferences can be made about the structure, function and nature of substrate binding of related members, based on the comparative analysis of similarities in proteins from other organisms.

2.2 Methods

2.2.1 Database searching: The amino acid sequences corresponding to mycolyl-transferases (Ag85A, Ag85B and Ag85C) from *M. tuberculosis* were obtained from the EBI (European Bioinformatics Institute) (http://srs.ebi.ac.uk/). Homologous proteins were identified for the species of Mycobacterium, Corynebacterium, and *N. farcinica* from the completed genome database using BLASTP and PSI-BLAST (Altschul *et al.,* 1990) (http://www.ncbi.nlm.nih.gov/BLAST/) with the Ag85B as the query sequence (Anderson *et al.,* 2001). The BLOSUM62 matrices were used and the results were sorted using E-value, with the gap costs set to existence at 11 and extension at 1.

2.2.2 Multiple sequence analysis: The multiple sequence alignment program CLUSTALW (Thompson *et al.,* 1994) available at EBI was used to align the mycolyl-transferases. The default parameters corresponding to a penalty of 10 for gap opening, 0.05 for gap extension and 8 for gap separation was assigned for the alignment.

2.2.3 Homology modeling: The 3-D models were constructed using the comparative modeling methods in the MODELLER program (Sali *et al.,* 1993) available in InsightII (version 2000.1, Accelrys Inc.) on a Silicon Graphics O2 Workstation (Silicon Graphics Inc) under UNIX operation system. The structures of Ag85A (PDB ID: 1SFR), Ag85B (PDB ID: 1F0P) and Ag85C (PDB ID: 1DQZ) were used as templates for modeling. Homology models were built for all the mycolyl-transferases from *N. farcinica* and corynebacterium species. The solvent accessibility of the protein active site was measured using the CASTp (Computed Atlas of Surface Topography of proteins) program which is an online resource that provides information for locating, delineating and measuring concave surface regions on 3-D structures of proteins. These include pockets located on surfaces and voids buried in the

interior of the proteins (Dundas *et al.,* 2006). It is available at the website (http://cast.engr.uic.edu).

2.2.4 Model evaluation: The models were evaluated using PROCHECK (Laskowski *et al.,* 1993). The RMSD values corresponding to the topologically equivalent residues in the structural superposition of models with crystal structures were derived using programs in InsightII. The method of Profiles-3D that measures the compatibility of an amino acid sequence with a protein of known 3-D structure was used to further assess the model (Lüthy, *et al.,* 1992).

2.2.5 Substrate docking: The trehalose substrate was docked into the binding site of all protein models using QUANTA 2000X, version 00.1110, Accelrys Inc. This enzyme-substrate complex was refined using molecular mechanics (MM) and molecular dynamics (MD) calculations to understand their interactions. Hydrogen atoms were added to the structures at pH 7.00 using BIOPOLYMER in InsightII. The default parameters, capping mode off was chosen such that the ends of the protein remain uncharged with NH2 and COOH groups. The forcefield, CVFF (Consistent Valence ForceField) was chosen, and the "Fix" option was used to select the potential atom types, partial charges and formal charges for the protein-substrate complex. The docked complex was subjected to energy minimization using 3000 steps steepest descent followed by conjugate gradient until an energy gradient of <0.01 kcal/mol/A^0 was achieved. The energy minimized structures were further subjected to MD simulations which were performed in the canonical ensemble (NVT) at 298^0 K using CVFF force field implemented in Discover3 (version 98.0) and equilibrated for 3000 femtoseconds with step size of 1 femtosecond.

2.2.6 Gene cluster analysis: The analysis of gene clusters was carried out by performing BLAST searches using mycolyl-transferases and their neighbouring proteins as query on all the finished and unfinished genomes. Mycolyl-

transferases and their flanking protein sequences (five or more on either side) were submitted to the BLASTP program (Altschul *et al.,* 1990) to identify sequence homologs.

2.2.7 Evolutionary Trace analysis: The evolutionary trace (ET) analysis was carried out using TraceSuite II server (Innis *et al.,* 2000) available at the website (http://www.cryst.bioc. cam.ac.uk/~jiye/evoltrace/evoltrace.html) by submitting the sequences corresponding to the carboxylesterase domain of CMN mycolyl-transferases and the crystal structure of Ag85B (PDB ID: 1F0P). A trace is generated by comparing the consensus sequences for groups of proteins that originate from a common node in a phylogenetic tree and is characterized by a common evolutionary time cut-off (ETC).

2.3 Results and Discussion

2.3.1 Comparative sequence analysis: Sequence searches identified varying number of mycolyl-transferases in each organism, four in *M. tuberculosis* and *C. diphtheria*, six in *C. glutamicum*, five in *C. efficiens* and thirteen in *N. farcinica*. The mycolyl-transferases corresponding to the mycobacteria species; *M. tuberculosis*, *M. leprae* and *M. bovis* are highly similar, therefore only the mycolyl-transferases from *M. tuberculosis* H37Rv strain are used in the subsequent discussions. Also, *M. tuberculosis* consists of a mycolyl-transferase precursor protein, MPT51 (GENE_ID: Rv3803) that does not possess mycolyl-transferase activity (Kremer *et al*., 2002; Wilson *et al*., 2004), therefore this sequence was not considered for subsequent analysis. The details of mycolyl-transferases analyzed and modeled in this work are provided in Table 2.1.

The multiple sequence alignment of the 31 mycolyl-transferases was generated using CLUSTALW and is shown in Figure 2.1. Inspite of the low sequence similarity shared between the proteins, we observed that 16 amino acid residues are conserved in all the sequences, these are L39, W51, P71, D81, W82, W97, F100, G124, S126, S150, D192, G214, E230, G260, H262 and W264 (amino acid numbering according to PDB ID: 1F0P). The alignment also indicated that some corynemycolyl-transferases and nocardiomycolyl-transferases have an insertion sequence of variable length (between 2 and 19 amino acid residues) that is located between the conserved G214 and E230. Further, two *N. farcinica* proteins, Nfa1810 and Nfa1820 consist of a 27 amino acid residue insertion sequence, that is rich in glycine and serine and present between the conserved W82 and W97, this can be seen from Figure 2.1. Generally it has been observed that the glycine and serine rich sequences are associated with cell-surface proteins.

Table 2.1. List of mycolyl-transferases in the CMN group of organisms.

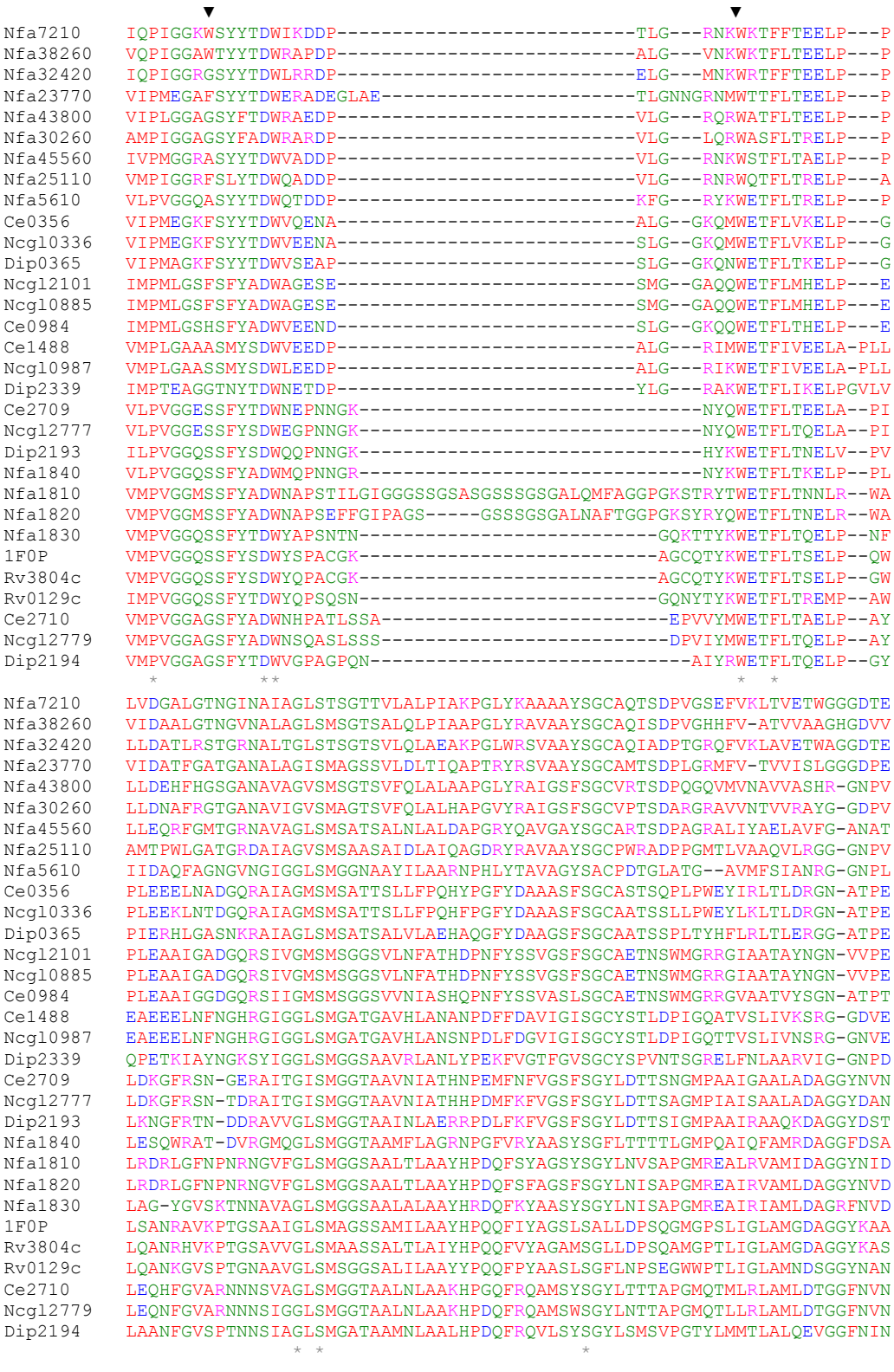
GENE_ID	GENBANK ID	Source	Length (aa)	Percentage Similarity with Ag85B	BLASTP E-value
Rv1886c	GI:15609023	*M. tuberculosis*	325	100	9e-173
Rv3804c	GI:15610940	*M. tuberculosis*	338	90	1e-146
Rv0129c	GI:57116693	*M. tuberculosis*	340	81	3e-123
Nfa1830	GI:54022147	*N. farcinica*	345	53	5e-48
Nfa1810	GI:54022145	*N. farcinica*	347	51	2e-47
Nfa1820	GI:54022146	*N. farcinica*	353	48	1e-45
Ncgl2777	GI:19554065	*C. glutamicum*	657	50	2e-44
Ce2709	GI:25029265	*C. efficiens*	669	52	5e-44
Nfa1840	GI:54022148	*N. farcinica*	624	50	1e-40
Ncgl2779	GI:19554067	*C. glutamicum*	341	50	2e-38
Dip2193	GI:38234734	*C. diphtheriae*	638	49	3e-38
Ce2710	GI:25029266	*C. efficiens*	360	51	9e-37
Dip2194	GI:38234735	*C. diphtheriae*	338	49	7e-35
Nfa5610	GI:54022528	*N. farcinica*	319	48	2e-33
Nfa30260	GI:54024995	*N. farcinica*	341	45	8e-28
Nfa32420	GI:54025211	*N. farcinica*	351	44	9e-27
Nfa38260	GI:54025796	*N. farcinica*	353	42	2e-26
Nfa7210	GI:54022688	*N. farcinica*	340	42	4e-26
Ncgl0987	GI:19552252	*C. glutamicum*	411	45	8e-26
Nfa25110	GI:54024480	*N. farcinica*	311	45	5e-25
Ce1488	GI:25028044	*C. efficiens*	390	43	9e-24
Dip0365	GI:38232981	*C. diphtheriae*	355	43	1e-23
Nfa45560	GI:54026529	*N. farcinica*	324	44	4e-23
Ncgl0885	GI:19552148	*C. glutamicum*	483	43	5e-23
Ncgl2101	GI:19553383	*C. glutamicum*	483	43	8e-23
Nfa23770	GI:54024346	*N. farcinica*	339	42	4e-22
Nfa43800	GI:54026351	*N. farcinica*	337	43	9e-22
Dip2339	GI:38234873	*C. diphtheriae*	406	44	3e-20
Ce0356	GI:25026912	*C. efficiens*	381	41	5e-20
Ce0984	GI:25027540	*C. efficiens*	484	42	1e-19
Ncgl0336	GI:19551592	*C. glutamicum*	365	42	8e-18

Figure 2.1: Multiple sequence alignment corresponding to CMN mycolyl-transferases.

```
Nfa7210     IKDDRNLRLYVYSAAMDENVIIDVQRPADASVPRPTLYLLNGAGGGEDDASWVAKSDALKFLSDKNVNV
Nfa38260    VVDARTVRLRVYSAAMGRVIDIDVQRPADTGAPRPTLYLLAGAGGGEDSASWAKQTSVLEFLADKNVNV
Nfa32420    AKEGRTWHLTVYSAAMDTEIAVDVQRPADDSVPAPNLYMLNGLDGGEGTASWAAATHALDWLADKPVNV
Nfa23770    GTPARLVDLAVYSPAMQRSIAVKVLRPADTTRPAPTLYLLNGAGGGEDAANWFGQTDAVEFFADKHVNV
Nfa43800    PENDRLLDLEIHSPAMDSTTRVLLLRAPDPDRPAPTLYLLNGASGHVDG-SWHDRTDYQRFFADKQVNV
Nfa30260    PRSDREVEVIVHSAAMAAEIPIRLLRAADPDRPAPTLYLLNGITGGGDGGNWFDRTGVAAFFAGEQVNV
Nfa45560    PLGGRQLEVVVHSAAMNRPITLWMS---HPGPGAPALYLLNAVDGGEDGGPWMNRTDVAAFFADKNVNV
Nfa25110    PLAPRVDQVQVYSPSMDAVVSSTVIR---ADGPAPTLYLLAGAGGGTDGISWWHHTDVRQFFADKNVNV
Nfa5610     ELSPTRSAVFVDSPAMGRVIQVQVLHP-AGGAARPSYYLLDGLDPGVGQSTWTNATDAEAFFRGKNVNV
Ce0356      ASGERVKEMWAYSPSMDRDVPLVVITADESAGPRPVIYLLNGGDGGEGNANWIMQTDVIDFYLEKNVNV
Ncgl0336    AADERVKEMWAYSPSMDRNVPLVVITADESAGPRPVIYLLNGGDGGEGAANWVMQTDVLDFYLEKNVNV
Dip0365     ATGDRVVEMWAHSPSMNRNVPLVVLKAANPG--RPTIYLLNGGDGGEGSANWVMQTKALDFYRDKDVNV
Ncgl2101    VDGDRIRQINAYSPSMGRTIPLVWVVPEDNTVPGPTVYALGGGDGGQGGQNWVTRTDLEELTSDNNINL
Ncgl0885    VDGDRIRQINAYSPSMGRTIPLVWVVPEDNTVPGPTVYALGGGDGGQGGQNWVTRTDLDELTSENNINL
Ce0984      VDGERIRQINAYSPSMERWIPLVWIVPEDTSEPRPTLYALGGGDGGQGSANWITKTDMPELMSSNNVHV
Ce1488      MDGLRLERWTVASPSMQRNVDVQIMRSVDAGAPAPMLYMLDGIGGNRNSSGWINHGQGPKVFGDENVTV
Ncgl0987    LNGLRLEKWSVASPSMQRNVDVQIMKSAEADSPAPMLYMLDGIGGNKNSSGWINGGEGPKVFADENVTV
Dip2339     DERFDVDRLFIESPAMRRIVQVQVQHPKDRTTPAPMLYLLDGVTAP-SQSGWLRKGDVQGAMANEHVTV
Ce2709      HVVLSIQSAAMPERPIKVQLLLPRDWYSSPDRDFPEIWALDGLRAIEKQSGWTIETNIEQFFADKNAIV
Ncgl2777    HVILTIQSAAMPERPIKVQLLLPRDWYSSPNREFPEIWALDGLRAIEEQSGWTIETNIEQYYADKNAIV
Dip2193     RVAVYVNTPSMG--QVQVQILLARDWFQDPNRSFPSVWALDGLRATDVENGWTIGTNIEQFYSDKNVNV
Nfa1840     RVALWVNSPSMG-APVQVQLLLARDWNAKPEARFPLLIMLDGLRATDDESGWTKDAGAEEFFADKNVTV
Nfa1810     SAAFNPDGFDFWVDSDMGPIKSRIFRA-ADGNTNRVVYALDGMRARNDLSGWEIDTEVARELTKWNINV
Nfa1820     SAAFDPAAFDFWVDSGMGPIKSRILRA-ADGNTNRVVYVLDGMRAPETLNGWEIETDVPALLASWNINV
Nfa1830     LRAPAGGYEELMVPSVMGPIKVQVQWA-SRG-GDAALYLLDGLRARDDRNAWSFETNAMEQFKNDNITL
1F0P        FSRPGLPVEYLQVPSPSMGRDIKVQFQ-SGGNNSPAVYLLDGLRAQDDYNGWDINTPAFEWYYQSGLSI
Rv3804c     FSRPGLPVEYLQVPSPSMGRDIKVQFQ-SGGANSPALYLLDGLRAQDDFSGWDINTPAFEWYDQSGLSV
Rv0129c     FSRPGLPVEYLQVPSASMGRDIKVQFQ-GGG--PHAVYLLDGLRAQDDYNGWDINTPAFEEYYQSGLSV
Ce2710      WDGVGYWVQRCDVYSPAMGRNIAVQIQPAQRGGNAGLYLLDGMRATTWSNAWLVDTNAAALYAPHNITL
Ncgl2779    WDAVGFWVQRCDVWSPAMGRNIPVQIQPAGRGGNAGLYLLDGMRATEYSNAWLVDTNAARLYAPNNITL
Dip2194     WDGVAHWVQRCDVFSPAMGRNITVQIQPAQRGGNAALYLLDGARANEIANAWTTDAHVQDLFVDHNITL
                                                    *           *
```

Conserved amino acid residues are (*), sites of insertion (▾).

Contd….


Nfa7210 IQPIGGKWSYYTDWIKDDP----------TLG---RNKWKTFFTEELP---P
Nfa38260 VQPIGGAWTYYTDWRAPDP----------ALG---VNKWKTFLTEELP---P
Nfa32420 IQPIGGRGSYYTDWLRRDP----------ELG---MNKWRTFFTEELP---P
Nfa23770 VIPMEGAFSYYTDWERADEGLAE----------TLGNNGRNMWTTFLTEELP---P
Nfa43800 VIPLGGAGSYFTDWRAEDP----------VLG---RQRWATFLTEELP---P
Nfa30260 AMPIGGAGSYFADWRARDP----------VLG---LQRWASFLTRELP---P
Nfa45560 IVPMGGRASYYTDWVADDP----------VLG---RNKWSTFLTAELP---P
Nfa25110 VMPIGGRFSLYTDWQADDP----------VLG---RNRWQTFLTRELP---A
Nfa5610 VLPVGGQASYYTDWQTDDP----------KFG---RYKWETFLTRELP---P
Ce0356 VIPMEGKFSYYTDWVQENA----------ALG--GKQMWETFLVKELP---G
Ncgl0336 VIPMEGKFSYYTDWVEENA----------SLG--GKQMWETFLVKELP---G
Dip0365 VIPMAGKFSYYTDWVSEAP----------SLG--GKQNWETFLTKELP---G
Ncgl2101 IMPMLGSFSFYADWAGESE----------SMG--GAQQWETFLMHELP---E
Ncgl0885 IMPMLGSFSFYADWAGESE----------SMG--GAQQWETFLMHELP---E
Ce0984 IMPMLGSHSFYADWVEEND----------SLG--GKQQWETFLTHELP---E
Ce1488 VMPLGAAASMYSDWVEEDP----------ALG---RIMWETFIVEELA-PLL
Ncgl0987 VMPLGAASSMYSDWLEEDP----------ALG---RIKWETFIVEELA-PLL
Dip2339 IMPTEAGGTNYTDWNETDP----------YLG---RAKWETFLIKELPGVLV
Ce2709 VLPVGGESSFYTDWNEPNNGK----------NYQWETFLTEELA--PI
Ncgl2777 VLPVGGESSFYSDWEGPNNGK----------NYQWETFLTQELA--PI
Dip2193 ILPVGGQSSFYSDWQQPNNGK----------HYKWETFLTNELV--PV
Nfa1840 VLPVGGQSSFYADWMQPNNGR----------NYKWETFLTKELP--PL
Nfa1810 VMPVGGMSSFYADWNAPSTILGIGGGSSGSASGSSSGSGALQMFAGGPGKSTRYTWETFLTNNLR--WA
Nfa1820 VMPVGGMSSFYADWNAPSEFFGIPAGS-----GSSSGSGALNAFTGGPGKSYRYQWETFLTNELR--WA
Nfa1830 VMPVGGQSSFYTDWYAPSNTN----------GQKTTYKWETFLTQELP--NF
1F0P VMPVGGQSSFYSDWYSPACGK----------AGCQTYKWETFLTSELP--QW
Rv3804c VMPVGGQSSFYSDWYQPACGK----------AGCQTYKWETFLTSELP--GW
Rv0129c IMPVGGQSSFYTDWYQPSQSN----------GQNYTYKWETFLTREMP--AW
Ce2710 VMPVGGAGSFYADWNHPATLSSA----------EPVVYMWETFLTAELP--AY
Ncgl2779 VMPVGGAGSFYADWNSQASLSSS----------DPVIYMWETFLTQELP--AY
Dip2194 VMPVGGAGSFYTDWVGPAGPQN----------AIYRWETFLTQELP--GY
LVDGALGTNGINAIAGLSTSGTTVLALPIAKPGLYKAAAAYSGCAQTSDPVGSEFVKLTVETWGGGDTE
VIDAALGTNGVNALAGLSMSGTSALQLPIAAPGLYRAVAAYSGCAQISDPVGHHFV-ATVVAAGHGDVV
LLDATLRSTGRNALTGLSTSGTSVLQLAEAKPGLWRSVAAYSGCAQIADPTGRQFVKLAVETWAGGDTE
VIDATFGATGANALAGISMAGSSVLDLTIQAPTRYRSVAAYSGCAMTSDPLGRMFV-TVVISLGGGDPE
LLDEHFHGSGANAVAGVSMSGTSVFQLALAAPGLYRAIGSFSGCVRTSDPQGQVMVNAVVASHR-GNPV
LLDNAFRGTGANAVIGVSMAGTSVFQLALHAPGVYRAIGSFSGCVPTSDARGRAVVNTVVRAYG-GDPV
LLEQRFGMTGRNAVAGLSMSATSALNLALDAPGRYQAVGAYSGCARTSDPAGRALIYAELAVFG-ANAT
AMTPWLGATGRDAIAGVSMSAASAIDLAIQAGDRYRAVAAYSGCPWRADPPGMTLVAAQVLRGG-GNPV
IIDAQFAGNGVNGIGGLSMGGNAAYILAARNPHLYTAVAGYSACPDTGLATG--AVMFSIANRG-GNPL
PLEEELNADGQRAIAGMSMSATTSLLFPQHYPGFYDAAASFSGCASTSQPLPWEYIRLTLDRGN-ATPE
PLEEKLNTDGQRAIAGMSMSATTSLLFPQHFPGFYDAAASFSGCAATSSLLPWEYLKLTLDRGN-ATPE
PIERHLGASNKRAIAGLSMSATSALVLAEHAQGFYDAAGSFSGCAATSSPLTYHFLRLTLERGG-ATPE
PLEAAIGADGQRSIVGMSMSGGSVLNFATHDPNFYSSVGSFSGCAETNSWMGRRGIAATAYNGN-VVPE
PLEAAIGADGQRSIVGMSMSGGSVLNFATHDPNFYSSVGSFSGCAETNSWMGRRGIAATAYNGN-VVPE
PLEAAIGGDGQRSIIGMSMSGGSVVNIASHQPNFYSSVASLSGCAETNSWMGRRGVAATVYSGN-ATPT
EAEEELNFNGHRGIGGLSMGATGAVHLANANPDFFDAVIGISGCYSTLDPIGQATVSLIVKSRG-GDVE
EAEEELNFNGHRGIGGLSMGATGAVHLANSNPDLFDGVIGISGCYSTLDPIGQTTVSLIVNSRG-GNVE
QPETKIAYNGKSYIGGLSMGGSAAVRLANLYPEKFVGTFGVSGCYSPVNTSGRELFNLAARVIG-GNPD
LDKGFRSN-GERAITGISMGGTAAVNIATHNPEMFNFVGSFSGYLDTTSNGMPAAIGAALADAGGYNVN
LDKGFRSN-TDRAITGISMGGTAAVNIATHHPDMFKFVGSFSGYLDTTSAGMPIAISAALADAGGYDAN
LKNGFRTN-DDRAVVGLSMGGTAAINLAERRPDLFKFVGSFSGYLDTTSIGMPAAIRAAQKDAGGYDST
LESQWRAT-DVRGMQGLSMGGTAAMFLAGRNPGFVRYAASYSGFLTTTTLGMPQAIQFAMRDAGGFDSA
LRDRLGFNPNRNGVFGLSMGGSAALTLAAYHPDQFSYAGSYSGYLNVSAPGMREALRVAMIDAGGYNID
LRDRLGFNPNRNGVFGLSMGGSAALTLAAYHPDQFSFAGSFSGYLNISAPGMREAIRVAMLDAGGYNVD
LAG-YGVSKTNNAVAGLSMGGSAALALAAYHRDQFKYAASYSGYLNISAPGMREAIRIAMLDAGRFNVD
LSANRAVKPTGSAAIGLSMAGSSAMILAAYHPQQFIYAGSLSALLDPSQGMGPSLIGLAMGDAGGYKAA
LQANRHVKPTGSAVVGLSMAASSALTLAIYHPQQFVYAGAMSGLLDPSQAMGPTLIGLAMGDAGGYKAS
LQANKGVSPTGNAAVGLSMSGGSALILAAYYPQQFPYAASLSGFLNPSEGWWPTLIGLAMNDSGGYNAN
LEQHFGVARNNNSVAGLSMGGTAALNLAAKHPGQFRQAMSYSGYLTTTAPGMQTMLRLAMLDTGGFNVN
LEQNFGVARNNNSIGGLSMGGTAALNLAAKHPDQFRQAMSWSGYLNTTAPGMQTLLRLAMLDTGGFNVN
LAANFGVSPTNNSIAGLSMGATAAMNLAALHPDQFRQVLSYSGYLSMSVPGTYLMMTLALQEVGGFNIN

```
Nfa7210     NMWGPPGSEEWVKNDPYVNAEGLRG---LELYISTGNGIPGPYDTLN-----GPYALPGSYGLANQILI
Nfa38260    NMYGPPDDPMWAANDPYVQAERLRG---LELFLSTGTGLPGKWDTLN-----GPHAMPGSDGLTNQLVL
Nfa32420    NMYGPDDSPLWRENDPVVNAEKLRG---TQLYISTGSGIPVLEDVQY-----YLNAAPGPMGAVN-LGL
Nfa23770    NMWGPTGGDGWREHDPYLQAHRLPP---IPMYISSGSGLPGPHDTLA-----NPRLHNDDRQLLNQTLV
Nfa43800    NMWGPPTDPTWRANDPYLHADRLRG---TAIYISSGSGLPGPLDNP-------AAVGGDPMQLGYQLLF
Nfa30260    NLWGPPEDPAWAANDPSLRAAELRD---TAVYVTAGTGRPGALDSLQ-----GPGIDADPLALADQLLI
Nfa45560    NMWGGPDSPLWAAHDPVLRAEELRG---LAIYVSAGDGRPGRHETLT-----APGIDGNPLDLVERTVV
Nfa25110    NMWGPPGDPGWQSHDAFRNAGALAG---KTVYLSAASGIPGPIDRGG-----LPAPT-----------
Nfa5610     NMWGPPGSPAWAEHDPARLAGNLRG---KTLYLSTGTGIPGPHEAEL-----KPQLAEN-------IFL
Ce0356      QMWGPRGGEVNIYNDALINSDKLRG---TDLYISNASGLAGHWESANSPRFNGLDQAYLSLAMTETIVT
Ncgl0336    QMWGPRGGEYNIYNDALINSDKLRG---TELYVSNASGLAGEWESVDSPRFEGLNQQVQSIAMAETVVT
Dip0365     QMWGPQGSEVNRYNDALINAERLRG---TEVYVSNNSGAVGKYDLPSSPRLAGKDPVTIFATNLITATE
Ncgl2101    QIFGEVDSDYSRYNDPLLNAAKLEE--QDNLYIFAGSGVFSELDVI------GDNAPIDEDAFKNRVLV
Ncgl0885    QIFGEVDSDYSRYNDPLLNAAKLEE--QDNLYIFAGSGVFSELDVI------GDNAPIDEDAFKNRVLV
Ce0984      QIFGEVDSDYARYNDPVINAHRLAK--QDNLYVFAASGVWSEVDVE------GENAPEDEKGLKNRITV
Ce1488      NMWGPVGSRTWQEHDVVSNPEGLRN---MAVYLSAANGVVDEIDREE-----------YADEPFYNLLA
Ncgl0987    NMWGPTGSETWKAHDVTSNPEGLRD---MAVYLSAANGVVDDIDLAD-----------SEKEPFYNLLA
Dip2339     LMWGRDITEQRRRNDVVANPSGIAS---MDTYIYVANGVATPSDVNG-----------PKEDGPFTLFG
Ce2709      AMWGPAGSERWLENDPKRNVDQLR--G-KQVYVSAGSGAD-DYGQDGSV------------ATGPANAA
Ncgl2777    AMWGPVGSERWQENDPKSNVDKLK--G-KTIYVSSGNGAD-DFGKEGSV------------AIGPANAA
Dip2193     AMWGPDGSQDWIDHDPKLGVEALR--G-ITTYVSAGSGRD-DFGEPGSV------------ANKKGSYA
Nfa1840     AMWGPPTSPEWEAHDPYLLADKLR--G-VSLYISSGSGTTGPFDQASGI------------PGVSTNYA
Nfa1810     AMAPPWG-PQWLRMDPFVFAPRLKANN-TRLWISAGSGLPGPADGFN-----------------FGTVN
Nfa1820     AMAPPWG-PQWLRMDPFVFAPNLIRNG-TRLWIAAASGLPTSTDPPS-----------------FNTLN
Nfa1830     SMAAPWS-PQWLRMDPFVFAPQLR--G-LPMYISAASGLPGQHDRPNSP------------VGVFNTGN
1F0P        DMWGPSSDPAWERNDPTQQIPKLVANN-TRLWVYCGNGTPNELGGAN---------------------IP
Rv3804c     DMWGPKEDPAWQRNDPLLNVGKLIANN-TRVWVYCGNGKPSDLGGNN---------------------LP
Rv0129c     SMWGPSSDPAWKRNDPMVQIPRLVANN-TRIWVYCGNGTPSDLGGDN---------------------IP
Ce2710      AMYGSVINPRRFENDPFWNMGGLR--G-KDVYVSAASGLWGPQDNGTR----------------VDHRIN
Ncgl2779    AMYGSIINPRRFENDPFWNMGGLA--N-TDVYISAASGLWSPQDDGVR----------------VDHRLT
Dip2194     NMYGSFFGLRRFQLDPLVNAAGLA--G-KDVYVSAASGIWGGPDYSYA----------------VNDRIN
                        *                       *

                ▼
Nfa7210     GGVIEAGTNYCTNNLKT--RLDELG-IPATYNFRPNGTHSWGYWNEEFPKSWPVLAKGL 291
Nfa38260    GGILEAGADHCTRNMRD--RLTQLG-IPATYDFQPRGTHSWGYWEDALKLSWPVLAKGL 290
Nfa32420    GVIIEAAVNQCTANLKN--RLDSLG-IPATYEFTPVGTHYWPYWEQALHDSWPMLAEGM 290
Nfa23770    GGAIESVTNLCTTRLAQ--RTAELGRTDITYNIRRPGTHSWGYWQDDLRDSWPMIARSI 298
Nfa43800    GAPLEAVTGMCTRQLRD--RLQELR-IPATVDLRPTGTHAWGYWQEDLHKAWPMFEAAL 287
Nfa30260    GGALEAVAADCTSELGA--RLRAAG-IPATVEVRPDGTHSWGYWEQDLRRCWPLFAAAL 290
Nfa45560    GGLMETVIGACTRPLVD--RLTSLA-VPATLALRP-GTHSWPYWQDDLHDSWPMFAAAI 286
Nfa25110    ---LEAIARTCTAAFAD--RLAELG-IAAVHVDRPLGAHTWGQFETDLHESWPHLAAAL 272
Nfa5610     GGPVEVGVNICTVAFEQ--RLRGLG-IPARIDYSPVGTHSWSYWQDTLHASWSTIGRAL 280
Ce0356      GGLIEAATNKCTHDLKA--KLDHAGIP-ADWNLRPTGTHSWGWWQDDLRGSWDTFARSF 296
Ncgl0336    GGIIEAATNKCTHDLKA--KLDSAGIP-ADWNLRPTGTHSWGWWQDDLRGSWTTFARAF 296
Dip0365     GGIIEAGTNMCTHDLKV--KMDSLNIP-ATFNFRNTGTHSWGYWEEDMVASWELFNMAF 294
Ncgl2101    GFEIEAMSNTCTHNLKA--ATDQMGIDNINYDFRPTGTHAWDYWNEALHRFFPLMMQGF 292
Ncgl0885    GFEIEAMSNTCTHNLKA--ATDQMGIDNINYDFRPTGTHAWDYWNEALHRFFPLMMQGF 292
Ce0984      GFRIEALSNTCTHNLKA--ATDYHGIDTIHYDFRPTGTHAWDYWNEALHRFFPLMMQGF 292
Ce1488      GTVLERGALSCTEALDDAMQD--AGMTHQVVDYKGAGAHNWRNFNEQLQPGWDAVKDAL 287
Ncgl0987    GVVLERGSLSCTEALDESMSR--AGMNHQVVDYKDSGTHNWRNFNPQLQPGWDAIKHAL 287
Dip2339     NIVLEKMSYRCTQELEASVREKIADPSRITFDYHDGGVHSWPYYRQQLPVAWANVSKGQ 289
Ce2709      GVGLELISRMTSQTFVD--AANGAG-VNVIANFRPSGVHAWPYWQFEMTQAWPYMADSL 282
Ncgl2777    GVGLEVISRMTSQTFVD--RASQAG-VEVVASFRPSGVHSWEYWQFEMTQAFPHIANAL 282
Dip2193     GIGLEVISRMTTETFVA--HARRAG-VEVQAFFRPSGVHDWPYWQFEMTQAWPYMANAL 280
Nfa1840     GTGLEILSRLTSQNFVT--KLGELQ-IPATVNYRASGTHSWPYWDFEMRQSWPQAAAAL 282
Nfa1810     AMGLEVLALANTRAFQV--RMATLGANNVTYDFPAVGVHNWRYWETEVYRMIPDLSANI 311
Nfa1820     GMGLEALALANTRAFQV--RMATLGGGNAVYSFPPFGIHAWNNWRDEAVRMMPDLSANI 306
Nfa1830     AMALEALSLVNTRAFQV--RLKSLG-IPAQFDFPATGTHSWKYWEGQLWNSRQGILDAL 284
1F0P        AEFLENFVRSSNLKFQD--AYNAAGGHNAVFNFPPNGTHSWEYWGAQLNAMKGDLQSSL 282
Rv3804c     AKFLEGFVRTSNIKFQD--AYNAGGGHNGVFDFPDSGTHSWEYWGAQLNAMKPDLQRAL 282
Rv0129c     AKFLEGLTLRTNQTFRD--TYAADGGRNGVFNFPPNGTHSWPYWNEQLVAMKADIQHVL 280
Ce2710      GSVLEAVSLATTRAWEA--KARAEG-LNVTADYPNTGIHSWAQFSSQLHKTRDRVLNVM 286
Ncgl2779    GSVLEFVAMTSTRIWEA--KARLQG-LNPTADYPMYGIHGWAQFNSQLERTQGRVLDVM 286
Dip2194     GSILEIASRVSTRIWEA--QARAIG-LNLTTNYPLLGVHNWVQWRYQIEQSKPRILDVM 283
               *                                 * * *
```

2.3.2 3-D modeling and structure analysis: Homology models were built for all the mycolyl-transferases from *N. farcinica*, *C. diphtheria*, *C. glutamicum* and *C. efficiencs* species. We have taken the crystal structures of Ag85B (Anderson *et al*., 2001; PDB ID: 1F0P) as template structure for generating 3-D models. The 3-D structure of all the mycolyl-transferases corresponds to a common α/β hydrolase fold (Figure 2.2a). The amino acid residues responsible for the mycolyl-transferase activity are S126, E230 and H262 in PDB ID: 1F0P. Evaluation of the 3-D models corresponding to corynemycolyl-transferases and nocardiomycolyl-transferases according to PROCHECK indicated more than 85% amino acid residues in the allowed regions of the Ramachandran plot (Ramachandran, *et al.,* 1963) suggesting that the models are of good quality. Further, according to the Profiles-3D, the 'observed' scores for the models lie between 124-134, close to the 'expected' scores suggesting the compatibility of sequence and structure. Also, the overall RMSD of the respective structures is ~0.68A^{o} and residues that form the active site S126, E230 and H262 are highly superimposed. The conservation of active site residues and their positions in the 3-D models indicated that all corynemycolyl-transferases and nocardio-mycolyl-transferases must also retain catalytic activity.

Examination of the models on computer graphics showed that the conserved residues L39, P71, D81, W82, W97 and F100 comprise the 'hydrophobic tunnel'. These are needed in order to accommodate the alkyl chain of mycolic acid, indicating a functional conservation in these proteins. The invariant S126 and G260 are close to the active site comprising E230. The indole side chains of W51 and W264 are perpendicular to each other and are in proximity to G124 associated with the β5 strand. The amino acid residue D192 is away from the active site indicating that the conservation extends beyond the active site in CMN mycolyl-transferases. We observed that the disulphide connectivity patterns are different in these proteins. The structures of 1SFR (Ag85A) and 1F0P (Ag85B) consist a disulphide bridge connecting half cystine

residues on β5 and β6 strands. In some proteins, half cystine residue on the α10 helix and half cystine residue on the loop connecting β6 strand and α6 helix are involved in the disulphide bridge. The information on the disulphide connectivity pattern is provided in Table 2.2. Based on the structural superposition, we observed that the differences between these structures are only in the loop regions. The 27 amino acid residue insertion in Nfa1810 and Nfa1820 is located between β5 and β6 strand and is away from the active site and we therefore predict that it may not interfere with the activity of the protein. According to the structure of 1F0P (Ag85B bound to the substrate trehalose), two substrate binding pockets are present. We observed that the variable region preceding the E230 forms an "insertion loop" close to the trehalose 1151 binding site (trehalose numbering according to 1F0P) (Figure 2.2b). The length and the amino acid composition of this insertion loop is variable and is given in Table 2.2. The protein with a long insertion loop formed a larger substrate binding pocket relative to the mycolyl-transferases. The corynemycolyl-transferases and nocardiomycolyl-transferases with large substrate binding pocket are: Nfa7210, Nfa38260, Nfa32420, Nfa23720 Nfa43800, Nfa30260, Nfa45560, Nfa25110, Nfa5610, Ce0356, Ncgl0336, Dip0365, Ncgl2101, Ncgl0885, and Ce0984.

Figure 2.2a: The α/β hydrolase fold of the mycolyl-tranferase Ag85B (PDB ID: 1F0P) (α helices are shown in red and β strands are shown in blue).

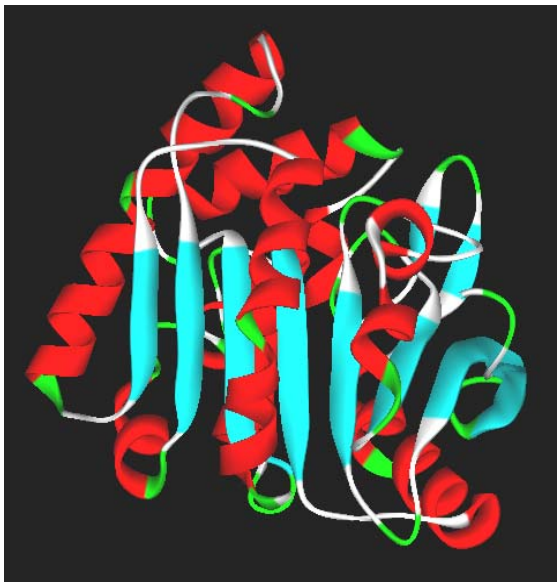


Figure 2.2b: The structural superposition of representative CMN mycolyl-transferases (PDB ID: 1F0P), Ncgl0336 (green), Ncgl0987 (pink). The side chains of the active site residues S126, E230, H262 (white) and trehalose 1151 (yellow) are represented in ball and stick model.

2.3.3 Docking analysis: In order to get an insight into the nature of interaction between the mycolyl-transferases and substrate-trehalose, trehalose was docked into the substrate binding site of all modeled structures after optimization using energy minimization. The specificity pockets defined by interactions with the trehalose substrate were examined and the results are presented in Table 2.2. While some proteins retain the nature of residues that line the specificity pockets, mutations such as D40N, R43D/G, S236N/A are observed in Nfa25110, Nfa45560, Nfa7210, Nfa38260, Nfa32420, Nfa23770, Nfa43800, Nfa30260, Dip0365, Ncgl0987, Ce1488, Ncgl0885, Ce0984, Ncgl2101, Ncgl0336 and Ce0356. In these proteins as a result of mutations, the substrate specificity may be affected. We observed that the proteins with specific amino acid mutations were associated with a large substrate binding site (see Figure 2.3). Also, the proteins comprising conserved amino acid residues in the substrate binding site are not associated with an insertion loop. We, therefore, infer that such proteins may bind trehalose.

It is often observed that, during the evolution, gene duplications, rearrangements and gene loss occurs in genomes due to a complex, general purpose mechanism for rapid adaptation of the organism. As a result of gene duplication, extra copies of selected genes are evolved. Duplications are important because they effectively allow at least one of the gene copies to evolve while the function of the original gene can remain intact. Many new functions arise from duplication and subsequent change of old genes. As a result, duplication of pre-existing genetic information provides the raw material from which new gene functions can evolve thereby contributing to the genetic complexity during evolution. With reference to mycolyl-transferases in the CMN genera, the presence of varying number of proteins in each organism reflects extensive gene duplication events during evolution of these organisms. Further, we identified that the overall structure, active site and hydrophobic tunnel are identical in all proteins, with significant differences in substrate

specificity pockets, which may be a result of selective pressure during evolution. From this work we propose that the trehalose is the original substrate and its binding is retained only in some corynemycolyl-transferases and nocardiomycolyl-transferases. During gene duplication, mutations in the substrate binding site have occurred such that the newly evolved proteins can bind to other sugars so as to synthesize organism specific polysaccharide-mycolate cell wall component.

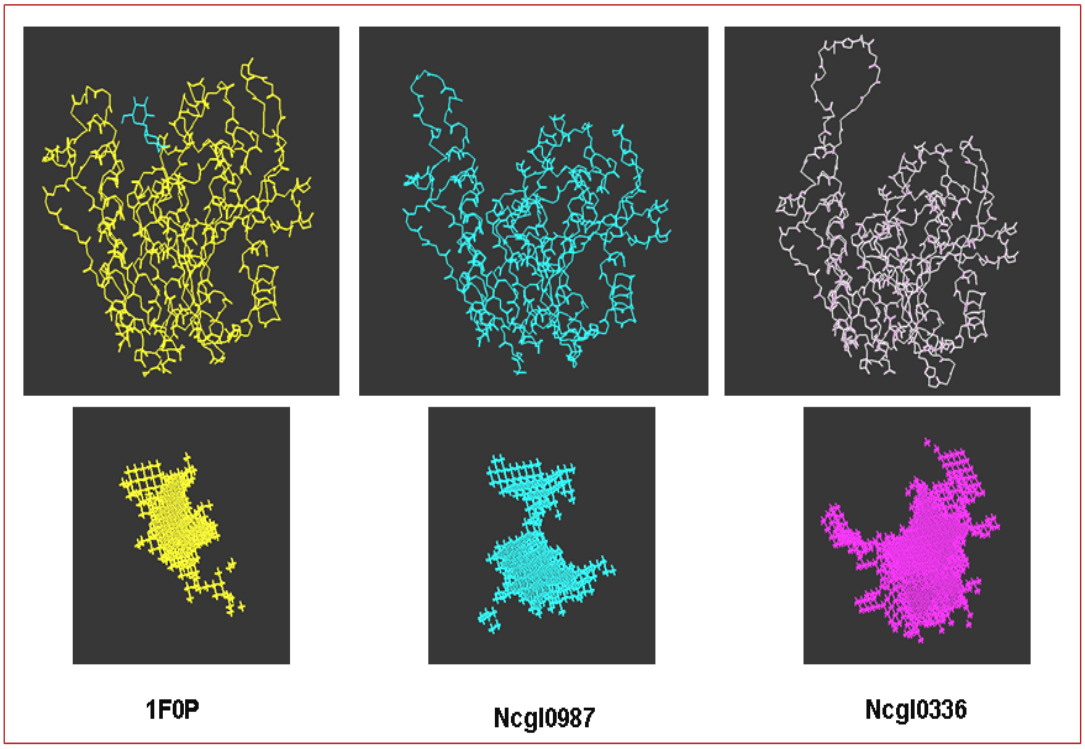
Table 2.2. Table showing 'Insertion loop' amino acid sequence, disulphide bridges and substrate binding pockets in CMN mycolyl-transferases.

Protein	'Insertion loop' amino acid sequence	Disulphide bridge	Trehalose 1151 binding residues						
1F0P	-	Cys 87-Cys 92	40D	43R	126S	223N	262H	263S	264W
Rv0129	-	---	38D	41R	124S	221N	260H	261S	262W
Rv3804	-	Cys 87-Cys 92	40D	43R	126S	223N	262H	263S	264W
Ncgl2777	AIGPA		40D	43R	121S	223N	261H	262S	263W
Ce2709	ATGPA		40D	43R	121S	223N	261H	262A	263W
Ncgl2779	DH		41D	44R	128S	228R	266H	267G	268W
Ce2710	DH		41D	44R	128S	228R	266H	267S	268W
Ncgl0987	SEKEPFYN		41D	44G	125S	221E	267H	268N	269W
Ce1488	YADEPFYN		41D	44G	125S	228L	267H	268N	269W
Ncgl0885	DNAPIDEDAFKNR		41G	44D	124S	219G	272H	273A	274W
Ce0984	ENAPEDEKGLKNR		41G	44D	124S	233I	272H	273A	274W
Ncgl2101	DNAPIDEDAFKNR		41G	44D	124S	219G	272H	273A	274W
Ncgl0336	SPRFEGLNQQVQSIAMAET		41N	44D	124S	218D	276H	277S	278W
Ce0356	SPRFNGLDQAYLSLAMTET		41N	44D	124S	229Y	276H	277S	278W
Nfa1810	FG		40D	43R	153S	252T	291H	292N	293W
Nfa1820	FN		40D	43R	148S	247T	286H	287A	288W
Nfa1830	SPVGVFN		39D	42R	124S	226T	264H	265S	266W
Nfa1840	PGVST		40L	43L	122S	225N	263H	264S	265W
Nfa25110	-	Cys 146-Cys 227	38A	41G	120S	214G	252H	253T	254W
Nfa45560	APGIDGNPLDLVER	Cys 146-Cys 242	40V	43G	122S	229T	266H	267S	268W
Nfa7210	GPYALPGSYGLANQ	Cys 149-Cys 246	43A	46G	123S	220P	271H	272S	273W
Nfa38260	GPHAMPGSDGLTNQ	Cys 150-Cys 246	42A	45G	124S	233L	271H	272S	273W
Nfa32420	YLNAAPGPMGAVN-	Cys 149-Cys 245	41N	44D	123S	232L	270H	271Y	272W
Nfa23770	NPRLHNDDRQLLNQ	Cys 156-Cys 252	41N	44G	130S	239T	278H	279S	280W
Nfa43800	AVGGDPMQLGYQ	Cys 148-Cys 242	41N	44S	122S	229L	267H	268A	269W
Nfa30260	GPGIDADPLALADQ	Cys 149-Cys 245	41N	44T	123S	219P	270H	271S	272W
Nfa5610	KPQLAEN	Cys 148-Cys 235	40D	43D	122S	222I	260H	261S	262W
Dip0365	SPRLAGKDPVTIFATNLIT		41G	44G	124S	220L	274H	275S	276W
Dip2339	PKEDGPFT		41D	44T	125S	228L	269H	270S	271W
Dip2193	ANKKG		40D	43R	121S	218A	261H	262D	263W
Dip2194	ND		41D	44R	125S	223N	263H	264N	265W

Contd…

Protein	'Insertion loop' amino acid sequence	Disulphide bridge	Trehalose 1152 binding residues						
1F0P	-	Cys 87-Cys 92	154D	157Q	159M	231N	232F	235S	236S
Rv0129	-	---	152N	155E	157W	229G	230L	233R	234T
Rv3804	-	Cys 87-Cys 92	154D	157Q	159M	231G	232F	235T	236S
Ncgl2777	AIGPA		149D	152S	154G	231V	232I	235M	236T
Ce2709	ATGPA		149D	152S	154G	231L	232I	235M	236T
Ncgl2779	DH		156N	159A	161G	236F	237V	240T	241S
Ce2710	DH		156T	159A	161G	236A	237V	240A	241T
Ncgl0987	SEKEPFYN		153S	156D	158I	236R	237G	240S	241C
Ce1488	YADEPFYN		153S	156D	158I	236R	237G	240S	241C
Ncgl0885	DNAPIDEDAFKNR		152E	154N	156W	241A	242M	245T	246C
Ce0984	ENAPEDEKGLKNR		152E	154N	156W	241A	242L	245T	246C
Ncgl2101	DNAPIDEDAFKNR		152E	154N	156W	241A	242M	245T	246C
Ncgl0336	SPRFEGLNQQVQSIAMAET		152A	155S	157L	246A	247A	250K	251C
Ce0356	SPRFNGLDQAYLSLAMTET		152S	155Q	157L	237T	238I	241G	242G
Nfa1810	FG		181N	184A	186G	260V	261L	264A	265N
Nfa1820	FN		176N	179A	181G	255A	256L	259A	260N
Nfa1830	SPVGVFN		152N	155A	157G	234A	235L	238V	239N
Nfa1840	PGVST		150T	153T	155G	233I	234L	237L	238T
Nfa25110	-	Cys 146-Cys 227	148W	151D	153P	222A	223I	226T	227C
Nfa45560	APGIDGNPLDLVER	Cys 146-Cys 242	150S	153A	155R	237T	238V	241A	242C
Nfa7210	GPYALPGSYGLANQ	Cys 149-Cys 246	151Q	154D	156V	241A	242G	245Y	246C
Nfa38260	GPHAMPGSDGLTNQ	Cys 150-Cys 246	152Q	155D	157V	241A	242G	245H	246C
Nfa32420	YLNAAPGPMGAVN-	Cys 149-Cys 245	151Q	154D	156T	240A	241A	244Q	245C
Nfa23770	NPRLHNDDRQLLNQ	Cys 156-Cys 252	158M	161D	163L	247S	248V	251L	252C
Nfa43800	AVGGDPMQLGYQ	Cys 148-Cys 242	150R	153D	155Q	237A	238V	241M	242C
Nfa30260	GPGIDADPLALADQ	Cys 149-Cys 245	151P	154D	156R	240A	241V	244D	245C
Nfa5610	KPQLAEN	Cys 148-Cys 235	150D	153L	155T	230V	231G	234I	235C
Dip0365	SPRLAGKDPVTIFATNLIT		152S	155L	157Y	244A	245G	248M	249C
Dip2339	PKEDGPFT		153S	156N	158S	236K	237M	240R	243Q
Dip2193	ANKKG		149D	152S	154G	231V	232I	235M	236T
Dip2194	ND		153S	156V	158G	233I	234A	237V	238S

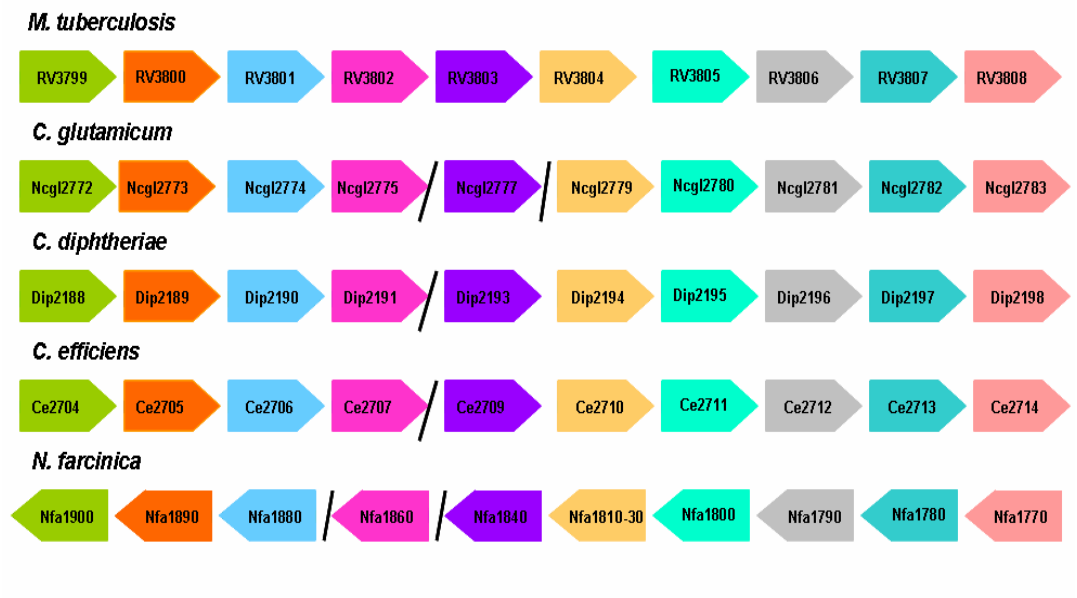
Figure 2.3: The proteins with large substrate binding pocket along with long insertion loops in (Ncgl0987 and Ncgl0336) are indicated with respect to the crystal structure (PDB ID:1F0P). Trehalose is indicated in blue in PDB ID:1F0P.



2.3.4 Gene cluster analysis: In order to establish the phylogenetic relationships between the mycolyl-transferases and to identify ancestral region among these proteins we have carried out BLASTP searches on various mycolyl-transferases and their flanking proteins. The analysis of all mycolyl-transferases and their neighbouring proteins revealed that genes between Rv3799–Rv3808 in the *M. tuberculosis* genome has corresponding orthologs in Corynebacterium and Nocardia genera and shown in Figure 2.4. The ten protein orthologs shown in Figure 2.4 share high sequence similarity in the five different species analyzed. In addition to mycolyl-transferase (Rv3804) and its precursor protein (Rv3803) this cluster also comprises propionyl CoA carboxylase (Rv3799), polyketide synthase (Rv3800), acyl CoA synthase (Rv3801), membrane proteins (Rv3806, Rv3807), and hypothetical proteins (Rv3802, Rv3805). We observed that the Nocardia proteins are arranged in the reverse order relative to the other species. We report that this set of genes represent the only mycolyl-transferase comprising gene cluster during divergence of a common ancestral organism into individual genera, such as, Corynebacterium, Mycobacterium and Nocardia (CMN group). Therefore, we propose that this gene cluster corresponds to the "Ancient Conserved Regions – ACR's" among the mycolyl-transferases across the CMN genera. It was reported that Rv3800 (*pks13*) is involved in the final condensation step in mycolic acid synthesis (Damien *et al.,* 2004). It was also reported that the genes; Rv3799, Rv3800 and Rv3801 (*accD4-pks13-fadD32*) play an essential role in the biosynthesis of mycolic acids (Gande *et al.,* 2004). These results indicate that the proteins in this cluster are important for the mycolic acid synthesis and its transfer to trehalose. Since, functionally related genes are often clustered, we suggest that the other "uncharacterized" proteins (Rv3802 and Rv3805) belonging to the ACR gene cluster may also have a role in associated functions. Further, we observed that the gene neighbours of mycolyl-transferase, Rv0129 and Rv1886 are conserved among *M. tuberculosis*

and *M. bovis* suggesting that gene duplication events have occurred before speciation.

Figure 2.4: Schematic representation of genes corresponding to the 'Ancient Conserved Regions' in five completed genomes based on gene neighbourhood analysis. '/' indicates insertion of gene. Nfa1810-30 indicates three genes Nfa1810, Nfa1820 and Nfa1830.



2.3.5 Evolutionary Trace analysis: The TraceSuite II server generates a phylogenetic tree split into 10 evenly distributed partitions (P01–P10) in the order of increasing evolutionary time cut-off (ETC) as shown in Figure 2.5a. The conserved amino acid residues associated with each partition is shown in Figure 2.5b. Analysis of amino acid residues corresponding to P01 partition (Figure 2.5b) revealed that 12 amino acid residues are "absolutely conserved". By examining the equivalent residues in the crystal structure of the protein (PDB ID: 1F0P), we infer that the residues; L39, P71, W82, W97 and F100 constitute the 'hydrophobic tunnel' as shown in Figure 2.6a. The residues in the 'hydrophobic tunnel' are needed in order to accommodate the alkyl chain of mycolic acid indicating a functional conservation in these proteins. According to Figure 2.5a, the 14 proteins indicated in the lower branch, from Corynebacterium, Mycobacterium and Nocardia represent the 'Ancient Conserved Region' proteins. The 18 proteins in the upper branch, correspond to Nocardia and Corynebacterium. From the multiple sequence alignment, we observed that the proteins in the upper branch of Figure 2.6a are associated with an insertion loop of variable length between 4 to 20 amino acid residues and this loop is close to the active site. The positions of the insertion loops in their 3-D structures are shown in Figure 2.6b. Further, the amino acid residues comprising the specificity pockets defined by interactions with trehalose substrate in the protein with PDB ID: 1F0P are mutated in these proteins. Primarily, the mutations associated with the substrate binding sites in some Corynebacterium (Adindla *et al.,* 2004a) and Nocardia proteins accompanied by the presence of 'insertion loops' close to the active site suggest that these may interfere with trehalose binding. These Corynebacterium and Nocardia proteins are possibly a result of divergent evolution accompanied by gene duplication and mutation events in order to accommodate different substrates in the binding site. This suggests that the ancient proteins form a distinct cluster and are different from proteins that evolved later.

The mycolyl-transferases Nfa1840, Ncgl2777, Ce2709 and Dip2193 comprise a 300 amino acid residue C-terminal extension as a result of gene fusion events. We previously reported that the corynemycolyl-transferase Ncgl2777 gene in *C. glutamicum* and Ce2709 gene in *C. efficiencs* (Adindla *et al.,* 2004b) are associated with a 55 amino acid residue 'LGFP' tandem repeats in the C-terminal region. Brand *et al.,* 2003 have demonstrated that the deletion of Ncgl2777 gene in *C. glutamicum* resulted in a 10-fold increase in cell volume of the organism thereby suggesting its involvement in cell shape formation. We suggested that the abnormal increase in the cell volume of *C. glutamicum* upon the deletion of the gene Ncgl2777 is due to the loss of C-terminal domain corresponding to the LGFP tandem repeats that may be responsible for maintaining the cell-wall integrity. In this work, we observed that the 'LGFP' tandem repeats are also present in the C-terminal region of Nocardia (Nfa1840) and *C. diphtheria* (Dip2193) proteins which imply that these proteins are also functional cell surface proteins and may be involved in maintaining cell wall integrity.

Figure 2.5a: TraceSuite II analysis representing partition based on evolutionary time cut-off.

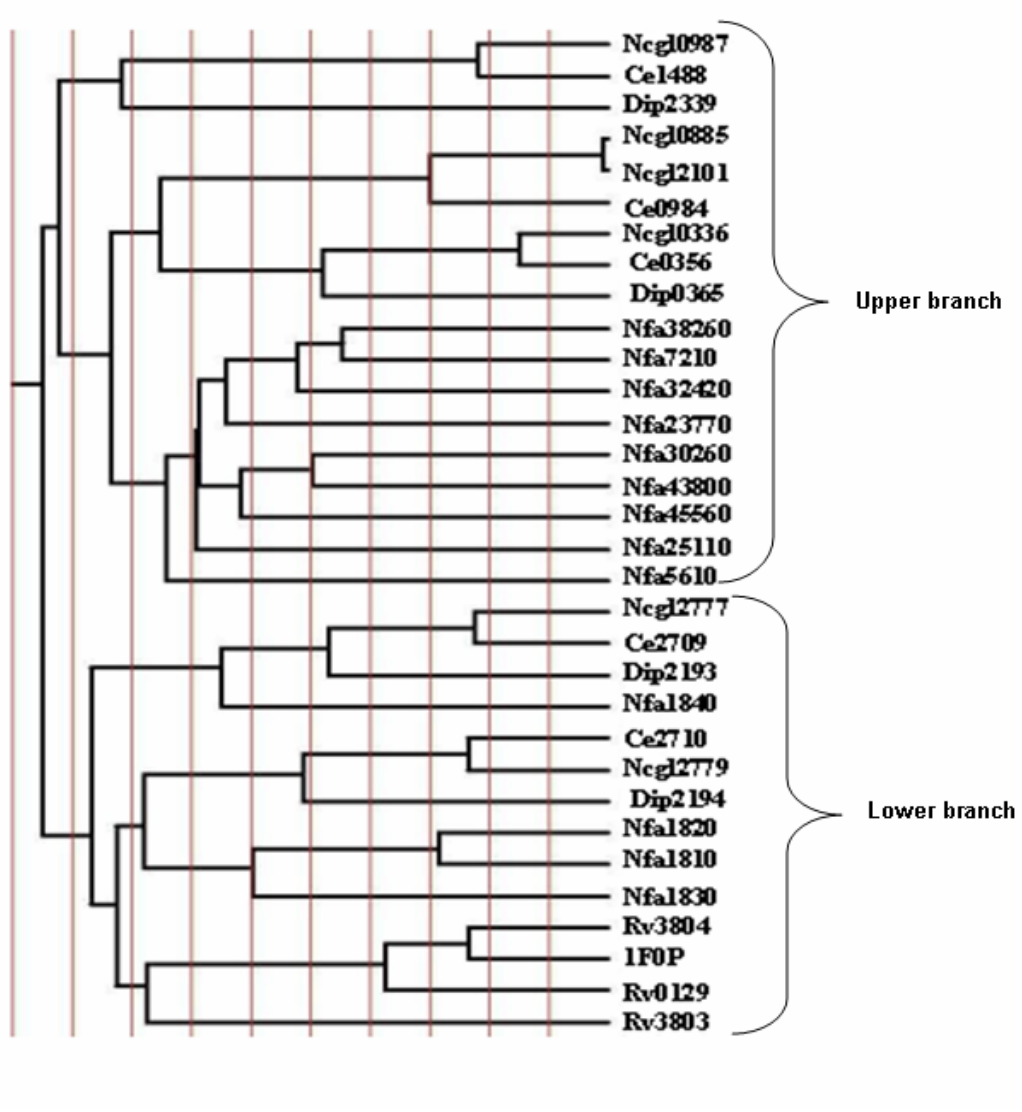
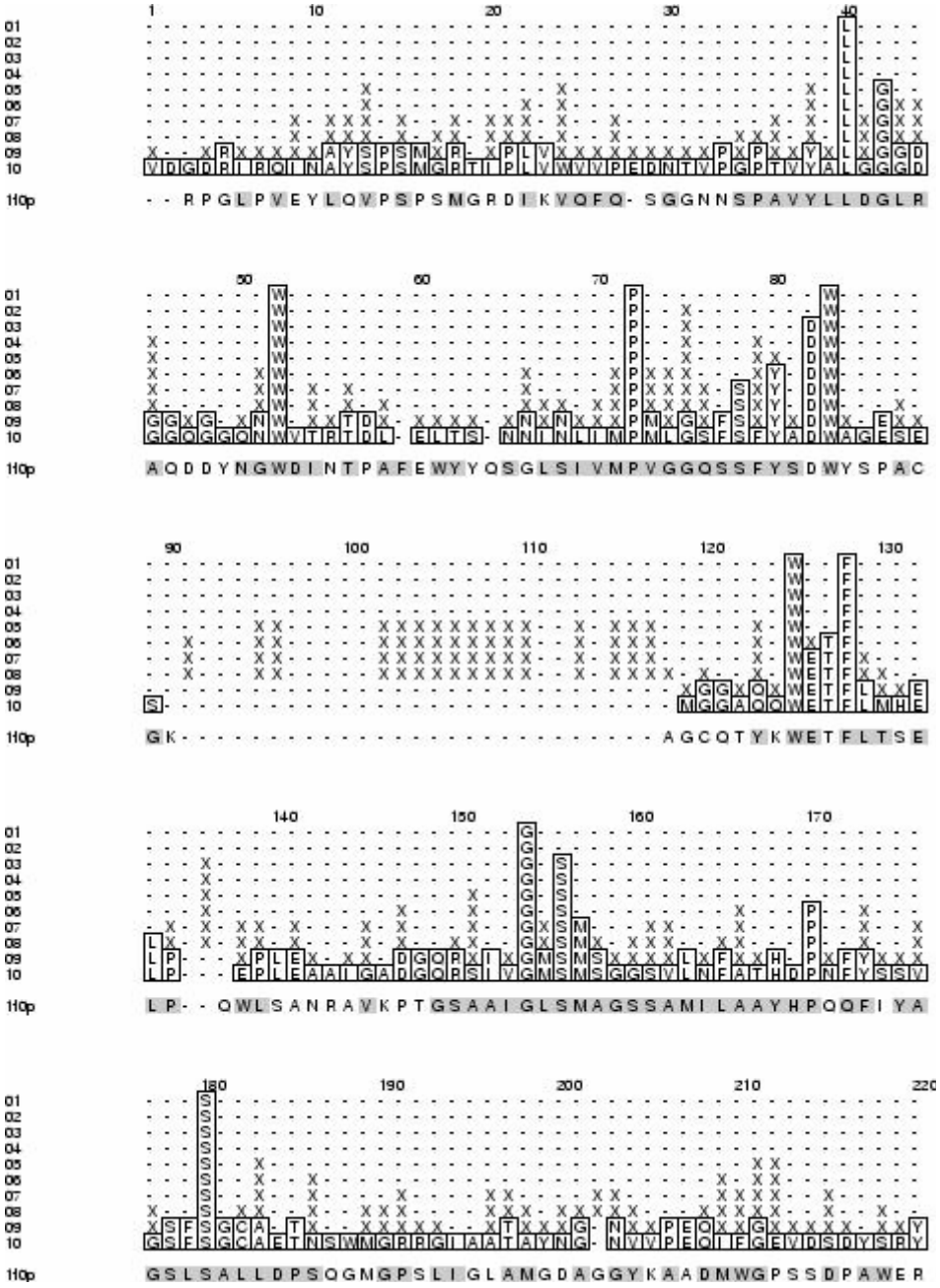


Figure 2.5b: 'Absolutely conserved' residues corresponding to P01 partition.



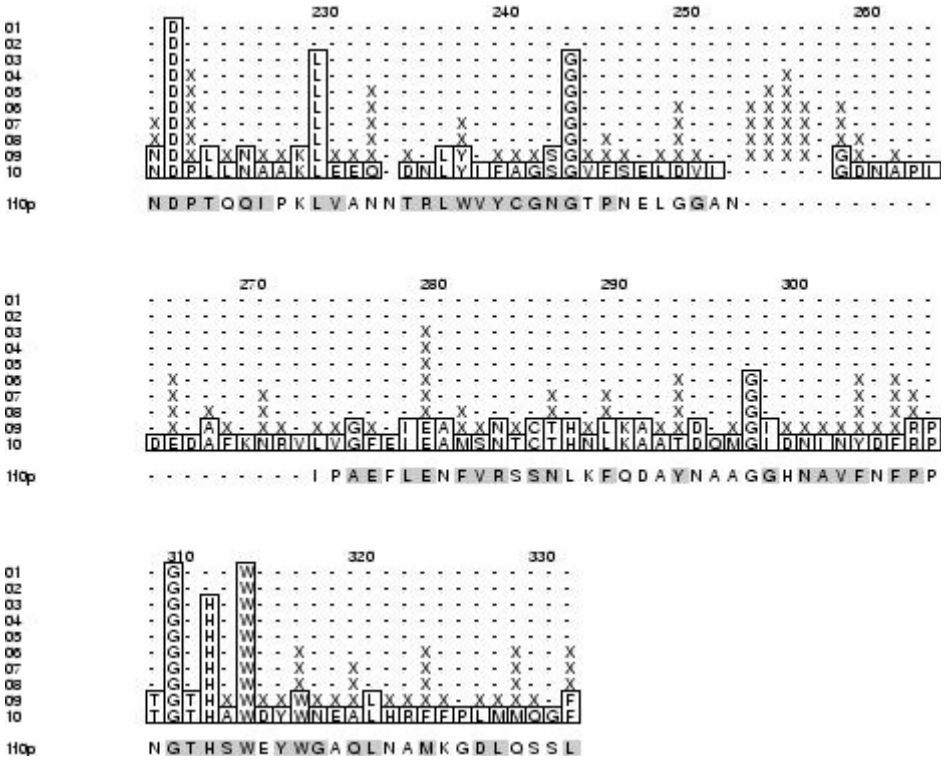

230
240
250
260
270
280
290
300
310
320
330
01
02
03
04
05
06
07
08
09
10
1Hqp
NDPTQQIPKLVANNTRLWVYCGNGTPNELGGAN
IPAEFLENFVRSSNLKFQDAYNAAGGHNAVFNFPP
NGTHSWEYWGAQLNAMKGDLQSSL

Figure 2.6a: Stereo-view showing 3-D model corresponding to the protein with GENE_ID Nfa1840 (pink). The amino acid residues comprising the catalytic triad (yellow), hydrophobic tunnel (blue) and trehalose (red) are also indicated.

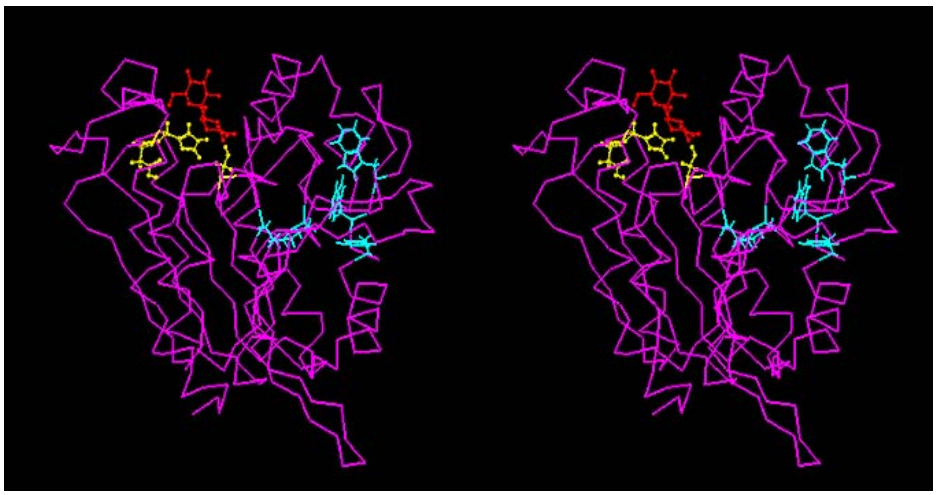
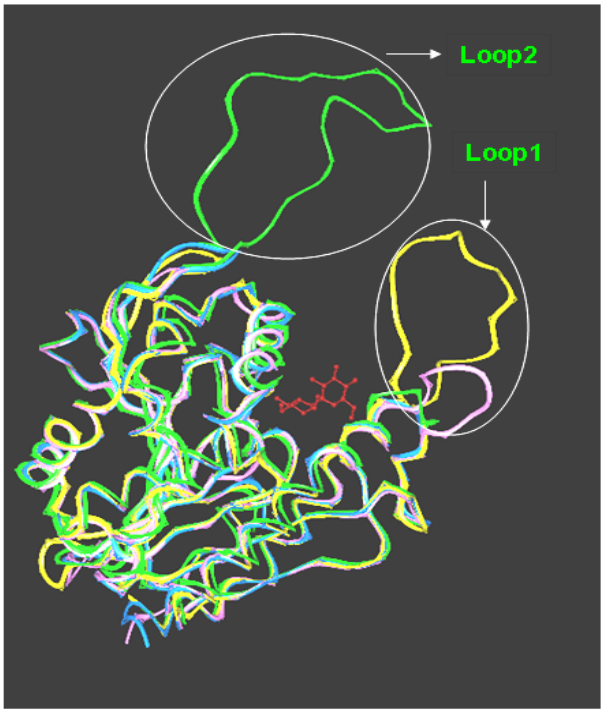


Figure 2.6b: Structural overlay corresponding to the proteins with PDB ID: 1F0P (blue) and GENE_IDs: Nfa1810 (green), Ncgl0336 (yellow) and Ncgl0987 (pink) indicating the position of the two loops; Loop 1 (D192 – E230 loop) and Loop 2 (W82-W97 loop).



2.4 Conclusions

1. We have identified, modeled and compared the 3-D structures of all the mycolyl-transerefases in the CMN genera. The overall α/β hydrolase fold characteristic of mycolyl-transerefases is conserved in all the proteins.

2. The two proteins of *N. farcinica*: Nfa1810 and Nfa1820, comprise a long insertion sequence of 27 and 22 amino acid residues rich in glycine and serine that is away from the active site and we predict that these may not be involved in the activity of the protein.

3. Based on the 3-D models, we propose that the proteins with long insertion loops Nfa25110, Nfa45560, Nfa7210, Nfa38260, Nfa32420, Nfa23770, Nfa43800, Nfa30260, Dip0365, Ncgl0987, Ce1488, Ncgl0885, Ce0984, Ncgl2101, Ncgl0336 and Ce0356 which have mutations in the key substrate binding pockets may not bind trehalose.

4. Based on gene cluster analysis, we have identified that the genes between Rv3799–Rv3808 in *M. tuberculosis* have orthologs in Corynebacteria, Mycobacteria and Nocardia (CMN) genomes. Therefore, this gene cluster possibly corresponds to the 'Ancient Conserved Region' of CMN mycolyl-transferases.

5. The evolutionary trace analysis suggests that 12 amino acid residues; L39, W51, P71, W82, W97, F100, G124, S126, D192, E230, G260 and W264 are 'absolutely conserved'. These amino acid residues constitute the active site and conserved hydrophobic tunnel in CMN mycolyl-transferases.

6. We observed that the LGFP tandem repeats are present in the C-terminal region of *N. farcinica* (Nfa1840) and *C. diphtheria* (Dip2193) proteins, which implies that these function as cell surface proteins and may be involved in maintaining the cell wall integrity.

2.5 References

Abou-Zeid, C., Ratliff, T. L., Wiker, H. G., Harboe, M., Bennedsen, J. & Rook, G. A. (1988). Characterisation of fibronectin-binding antigens released by *Mycobacterium tuberculosis* and *Mycobacterium bovis* BCG. *Infect. Immun*. **56,** 3046-3051.

Adindla, S., Guruprasad, K. & Guruprasad, L. (2004a). Three-dimensional models and structure analysis of corynemycolyltransferases in *Corynebacterium glutamicum* and *Corynebacterium efficiens*. *Int. J. Biol. Macromol*. **34**, 181-189.

Adindla, S., Inampudi, K. K., Guruprasad, K. & Guruprasad, L. (2004b). Identification and analysis of novel repeats in the cell surface proteins of archaeal and bacterial genomes using computational tools. *Comp. Funct. Genom*. **5,** 2–16.

Alashamaony, L., Goodfellow, M. & Minnikin, D. E. (1976). Free mycolic acids as criteria in the classification of Nocardia and the 'rhodochrous' complex. *J. Gen. Microbiol.* **92,** 188-199.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. (1990). Basic local alignment search tool. *J. Mol. Biol*. **215**, 403-410.

Anderson, D. H., Harth, G., Horwtiz, M. A. & Eisenberg, D. (2001). An interfacial mechanism and a class of inhibitors inferred from two crystal structures of the *Mycobacterium tuberculosis* 30 kDa Major secretory protein (antigen 85B), a mycolyl transferase. *J. Mol. Biol*. **307**, 671-681.

Belisle, J. T., Vissa, V. D., Sievert, T. Takayama, K., Brennan, P. J. & Besra, G. S. (1997). Role of the major antigen of *Mycobacterium tuberculosis* in the cell wall biogenesis. *Science,* **276**, 1420–1422.

Brand, S., Niehaus, K., Puhler, A. & Kalinowski, J. (2003). Identification and functional analysis of six mycolyltransferase genes of Corynebacterium glutamicum ATCC 13032: the genes cop1, cmt1, and cmt2 can replace each other in the synthesis of trehalose dicorynomycolate, a component of the mycolic acid layer of the cell envelope. *Arch. Microbiol*. **180**, 33-44.

Brennan, P. J. & Nikaido, H. (1995). The envelope of mycobacteria. *Annu. Rev. Biochem*. **96**, 29-63.

Cerdeno-Tarraga, A. M., Efstratiou, A., Dover, L. G., Holden, M. T., Pallen, M., Bentley, S. D., Besra, G. S. *et al*. (2003). The complete genome sequence and analysis of Corynebacterium diphtheriae NCTC13129. *Nucl. Acids Res.* **31,** 6516-6523.

Cocito, A. & Delville, J. (1985). Biological, chemical, immunological and staining properties of bacteria isolated from tissues of leprosy patients. *Eur. J. Epidemiol.* **1,** 202–231.

Cole, S. T., Brosch, R., Parkhill, J., Garnier, T., Churcher, C., Harris, D., Gordon, S. V. *et al*. (1998). Deciphering the biology of Mycobacterium tuberculosis from the complete genome sequence. *Nature*, **393**, 537-544.

Collins, M. D., Goodfellow, M. & Minnikin, D. E. (1982). A survey of the structures of mycolic acids in Corynebacterium and related taxa. *J. Gen. Microbiol.* **128**, 129-149.

Daffé, M. & Draper, P. (1998). The envelope layers of mycobacteria with reference to their pathogenicity. *Adv. Microb. Phys.* **39**, 131-203.

Damien, P., De Sousa-D'Auria, C., Houssin , C., Grimaldi , C., Chami, M. , Daffé, M. & Guilhot, C. (2004). A polyketide synthase catalyzes the last condensation step of mycolic acid biosynthesis in mycobacteria and related organisms. *Proc. Natl. Acad. Sci. USA*, **101**, 314-319.

Dandekar, T., Snel, B., Huynen, M. & Bork, P. (1998). Conservation of gene order: a finger print of proteins that physically interact. *Trends Biochem. Sci.* **23**, 324-328.

De Sousa-D' Auria, C., Kacem, R., Puech, V., Tropis, M., Leblon, G., Houssin, C. & Daffe, M. (2003). New insights into the biogenesis of the cell envelope of corynebacteria: identification and functional characterization of five new mycolyltransferase genes in *Corynebacterium glutamicum*. *FEMS Micrbiol. Letters*, **224,** 35-44.

Doolittle, W. F. (1998). You are what you eat: A gene transfer ratchet could account for bacterial genes in eukaryotic nuclear genomes. *Trends Genet.* **14,** 307–311.

Doolittle, W. F. (1999). Lateral genomics. *Trends Cell Biol.* **9,** M5–8.

Doolittle, R. F. (1999). Do you did my groove? *Nat. Genet.* **23**, 6-8.

Doolittle, W. F. (2000). Uprooting the tree of life. *Sci. Am.* **282,** 90–95.

Draper, P. (1998). The outer parts of the mycobacterial envelope as permeability barriers. *Frontiers Biosci.* **3**, d1253-d1261.

Dundas, J., Ouyang, Z., Tseng, J., Binkowski, A., Turpaz, Y. & Liang, J. (2006). CASTp: computed atlas of surface topography of proteins with structural and topographical mapping of functionally annotated residues. *Nucl. Acids Res.* **34,** W116-W118.

Enright, A., Ilipoulos, I., Kyrpides, N. & Ouzounis, C. (1999). Protein interaction maps for complete genomes based on gene fusion events. *Nature*, **402**, 86-90.

Gande, R., Gibson, K. J. C., Brown, A. K., Krumbach, K., Dover, L. G., Sahm, H., Shioyama, S. *et al*. (2004). Acyl-CoA Carboxylases (*accD2* and *accD3*), Together with a Unique Polyketide Synthase (*Cg-pks*), Are Key to Mycolic Acid Biosynthesis in *Corynebacterianeae* Such as *Corynebacterium glutamicum* and *Mycobacterium tuberculosis*. *J. Biol. Chem*. **279**, 44847-44857.

Glickman, M. S., Cox, J. S. & Jacobs Jr, W. R. (2000). A novel mycolic acid cyclopropane synthetase is required for cording, persistence, and virulence of *Mycobacterium tuberculosis*. *Mol. Cell.* **5**, 717-727.

Huynen, M. A. & Snel, B. (2000). Gene and context: Integrative approaches to genome analysis. *Advan. Protein Chem.* **54**, 345-379.

Innis, C. A., Shi, J. & Blundell, T. L. (2000). Evolutionary Trace Server (TraceSuite II). *Protein Eng.* **13,** 839-847.

Ishikawa, J., Yamashita, A., Mikami, Hoshino, Y., Kurita, H., Hotta, K., Shiba, T. *et al*. (2004). The complete genomic sequence of *Nocardia farcinica* IFM 10152. *Proc. Natl. Acad. Sc*i. USA, **101**, 14925–14930.

Kalinowski, J., Bathe, B., Bartels, D., Bischoff, N., Bott, M., Burkovski, A., Dusch, N. *et al*. (2003). The complete Corynebacterium glutamicum ATCC 13032 genome sequence and its impact on the production of L-aspartate-derived amino acids and vitamins. *J. Biotechnol.* **104,** 5-25.

Katsube, T., Matsumoto, S., Takatsuka, M., Okuyama, M., Ozeki, Y., Naito, M., Nishiuchi, Y. *et al.* (2007). Control of cell wall assembly by a histone-like protein in Mycobacteria. *J. Bacteriol.* **22,** 8241-8249.

Kawarabayasi, Y., Yamazaki, J., Hino, Y., Kikuchi, H., Nakamura, Y., Ikeo, K., Suzuki, M. *et al*. (2002). The entire genomic sequence of *Corynebacterium efficiens* YS-314. Unpublished.

Koonin, E. V., Makarova, K. S. & Aravind, L. (2001). Horizontal gene transfer in prokaryotes: quantification and classification. *Annu. Rev. Microbiol.* **55,** 709–742.

Koonin, E. V. & Galperin, M. Y. (2002). *Sequence—Evolution—Function. Computational Approaches in Comparative Genomics*. New York: Kluwer.

Koonin, E. V. (2005). Orthologs, Paralogs, and Evolutionary Genomics. *Annu. Rev. Genet.* **39,** 309–338.

Kremer, L., Maughan, W. N., Wilson, R. A., Dover, L. G. & Besra, G. S. (2002). The *M. tuberculosis* antigen 85 complex and mycolyltransferase activity. *Lett. Appl. Microbiol*. **34**, 233-237.

Kunin, V. & Ouzounis C. A. (2003). The balance of driving forces during genome evolution in prokaryotes. *Genome Res.* **13,** 1589–1594.

Laskowski, R. A., Mac Arthur, M. W., Moss, D. S. & Thornton, J. M. (1993). PROCHECK: a program to check the stereo chemical quality of protein structures. *J. Appl. Crystallog.* **26,** 283-291.

Lawrence (1997). Selfish operons and speciation by gene transfer. *Trends Micribiol*. **5**, 355-359.

Lawrence, J. G. & Hendrickson, H. (2003). Lateral gene transfer: When will adolescence end? *Mol. Microbiol.* **50,** 725–727.

Lüthy, R., Bowie, J. U. & Eisenberg, D. (1992). Assessment of protein models with three-dimensional profiles. *Nature,* **356**, 83-85.

Marcotte, E. M., Pellegrini, M., Ng, H., Rice, W. D., Yeates, T. O. & Eisenberg, D. (1999). Detecting protein function and protein-protein interactions from genome sequences. *Science*, **285**, 751-753.

Minnikin, D. E., Goodfellow, M. & Collins, M. D. (1978). Coryneform bacteria (Bousfield, I. J. & Callely, A. G. eds), pp. 85-160, Academic Press, London.

Mirkin, B. G., Fenner, T. I., Galperin, M. Y. & Koonin, E. V. (2003). Algorithms for computing parsimonious evolutionary scenarios for genome evolution, the last universal common ancestor and dominance of horizontal gene transfer in the evolution of prokaryotes. *BMC Evol. Biol.* **3,** 2.

Overbeek, R., Fonstein, M., D'Souza, M., Pusch, G. D. & Maltsey, N. (1999). The use of gene clusters to infer functional coupling. *Proc. Natl. Acad. Sci. USA*, **96**, 2896-2901.

Pennisi, E. (1998). Genome data shake tree of life. *Science,* **280,** 672–674.

Pennisi, E. (2001). Microbial genomes. Sequences reveal borrowed genes. *Science,* **294,** 1634–1635.

Ramachandran, G. N., Ramakrishnan, C. & Sasisekharan, V. (1963). Stereochemistry of polypeptide chain configurations. *J. Mol. Biol.* **7**, 95-99.

Ratliff, T. L., Mc Garr, J. A., Abou-Zeid, C., Rook, G. A., Stanford, J. L., Aslanzadeh, J. & Brown, E. J. (1988). Attachment of mycobacteria to fibronectin-coated surfaces. *J. Gen. Microbiol*. **134**, 1307-1313.

Ronning, D. R., Klabunde, T., Besra, G. S., Vissa, V. D., Belisle, J. T. & Sacchettini, J. C. (2000). Crystal structure of the secreted form of antigen 85C reveals potential targets mycobacterial drugs and vaccines. *Nature Struct. Biol*. **7**, 141-146.

Ronning, D. R., Vissa, V., Gurdyal, B., Belisle, J. T. & Saccchettini, J. C. (2004). Mycobacterium tuberculosis Antigen 85A and 85C structures confirm binding orientation and conserved substrate specificity. *J. Biol. Chem.* **279**, 36771-36777.

Sali, A. & Blundell, T. L. (1993). Comparative protein modeling by satisfaction of spatial restraints. *J. Mol. Biol*. **234**, 779-815.

Snel, B., Bork, P. & Huynen, M. A. (2002). Genomes in flux: the evolution of archaeal and proteobacterial gene content. *Genome Res.* **12,** 17–25.

Tabbara, K. F. (2007). Tuberculosis. *Curr. Opin. Ophthalmol.* **18,** 493-501.

Thompson, J. D., Higgins, D. G. & Gibson, T. J. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucl. Acids Res.* **22**, 4673-4680.

Wapinski, I., Pfeffer, A., Friedman, N. & Regev, A. (2007). Automatic genome-wide reconstruction of phylogenetic gene trees. *BIOINFORMATICS*, **23**, i549–i558.

Wiker, H. G. & Harboe, M. (1992). The antigen 85 complex: a major secretion product of *Mycobacterium tuberculosis*. *Microbiol Rev*. **56**, 648-661.

Wilson, R. A., Maughan, W. N., Kremer, L., Besra, G. S. & Futterer, K. (2004). The structure of Mycobacterium tuberculosis MPT51 (FbpC1) defines a new family of non-catalytic alpha/beta hydrolases. *J. Mol. Biol*. **335**, 519-530.

CHAPTER 3

In Silico Method for the Automated Identification of Novel Repeats in Complete Proteomes

3.1 Introduction

Biological sequence repeats are arranged in tandem patterns and are widespread in DNA and proteins. Proper delineation of repeats at the sequence level is not only important for understanding the structure and function of proteins, but is crucial for the detection of homologous sequences and to explain their evolutionary lineage.

The repeats and domains as discussed in Chapter 1 are characterized by conserved sequence motifs that may be identified according to the conservation of individual amino acid residues at equivalent positions derived from multiple sequence alignments. Repeats may be identified by manual examination, if the sequence similarity is very high and present in tandem. Programs such as BLASTP (Altschul *et al.,* 1990) are also useful in detecting internal and homologous repeats in a protein database. Several web based methods are available for *ab initio* identification of sequence repeats in proteins. Examples are RADAR (Heger & Holm, 2000), REP Program (Andrade *et al.,* 2000), REPRO (Heringa & Argos, 1993), PROSPERO (Mott, 2000) and TRUST (Szklarczyk & Heringa, 2004). These methods are described in detail in Chapter 1.

In our work, we have implemented TRUST as the main program for repeat identification method. We have downloaded and installed TRUST on the local Pentium IV computers on the Linux platform. TRUST program (Tracking Repeats Using Significance and Transitivity) (Szklarczyk & Heringa, 2004) exploits the concept of transitivity of alignments as well as a statistical scheme optimized for the evaluation of repeat significance. It detects repeats using the Waterman-Eggert algorithm (Waterman & Eggert, 1987). Transitivity is employed as the key strategy to assess the statistical significance (p-value) of repeat alignment scores, as opposed to various parameters and arbitrary thresholds used by other methods. It uses logical inference from alignments for

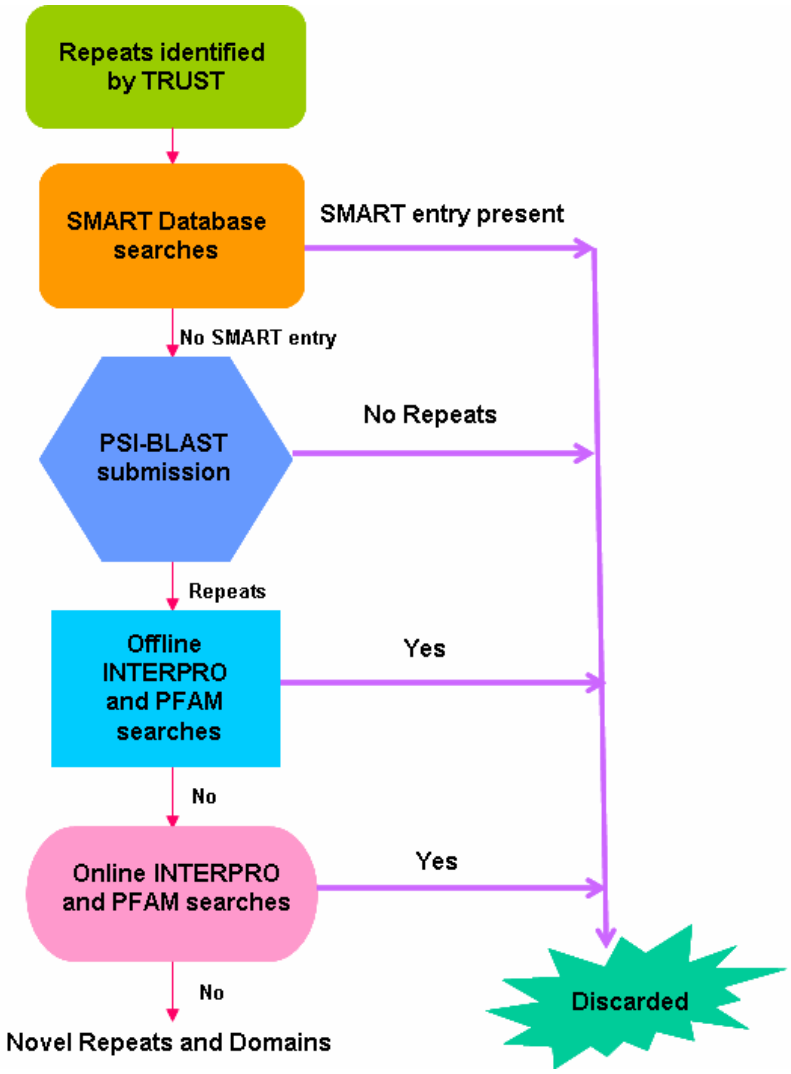
which the information exists that identifies distant homologous regions and at the same time can support or contradict existing sub-optimal alignments. The transitivity scheme enables to accurately calculate the repeat length and allows the generation of virtually noise free and sensitive profiles.

TRUST server together with the source code is available at (http://ibivu.cs.vu.nl/programs/trustwww).

3.2 Methods

The various steps used to carried out for the novel repeat identification in complete proteomes in this work is shown in Figure 3.1 and we discuss them below.

Figure 3.1: Systematic analysis of novel repeat identification in complete proteome shown in flow chart.



3.2.1 Download the complete organism proteome: We have downloaded each individual proteome, for example (bacterial Ex. *Bacillus anthracis* str. Ames), archaeal (13 organisms from archaeal origin) and the human proteome from the NCBI website in FASTA format.

3.2.2 Separate the proteome into individual files: All the protein sequences in the complete proteome were directly submitted to TRUST using sh findrepeats.sh command when the file size was less than 2MB. The files that are 2MB or larger in size, were divided into two or three parts such that each part is equal to or less than 1.5 MB by using, split –l command on Linux. This is because, TRUST cannot process files that are greater than 2MB in size. We have divided the human proteome using a PERL script "Sep_*protein.pl*" which separates all protein sequences into individual files and generates them in a FASTA format. The script is as follows.

Seq_protein.pl

```
# Open Human Proteome File
open ( File_data,"<Human_prot.dat");
# Read Human Proteome file
while (<File_data>)
{
# Check line start with '>'
        if ($_=/^>/)
        {
                ($Temp, $Temp_data) = split (/^/, $_);
# Get Protein sequence File name
                 ($File_name, $Temp) = split (/\|/, $Temp_data);
 # Create Protein sequence File
                open (Write_file,">$File_name");
        }
# Write data in Newly created Protein sequence File
        print write $_;
}
# Close Human Proteome File
close (File_data);
```

3.2.3 Identify the repeats in each protein sequence using TRUST: The separated protein files in FASTA format for example, GENOME_ID_part1.faa and GENOME_ID_part2.faa with .faa (FASTA) as extension were submitted to TRUST using the shell script findrepeats.sh command as long as the total size of the file does not exceed 2MB (GENOME_ID is assigned by NCBI and is unique to each organim). The command to submit is sh findrepeats.sh. Similarly, we can submit multiple sequences in multiple sessions of the same program. Our computer configuration supported submission of the same program at a time with commands sh findrepeats2.sh, sh findrepeats3.sh, sh findrepeats4.sh and sh findrepeats5.sh commands.

The script is as follows:

```
#!/bin/sh
# Declare Array
declare -a arr1
declare -a arr2
declare -a arr3
declare -a arr4
declare -a arr5
declare -a array

# Get list of all Directory in Array 'arr1'
arr1=`ls`
for i in $arr1
        do
                if [ $i ]
                        then
# Go in child Directory
                        cd $i

# Get list of Directory and file in child directory
                        arr2=`ls`

                        for j in $arr2
                        do
                        if [ $j ]
                        then
```

```
# Create Directory for each file
                    mkdir $j.dir
# Move File to respectve newly created Directory
                   mv $j $j. dir
                    cd $j.dir
# Create 'reapeats', 'all_sequences', and 'rep_sequences'for each file Directories
                    mkdir repeats
                    mkdir all_sequences
                    mkdir rep_sequences
# Spliting of sequence
                   seqretsplit -auto -sprotein1 -sformat1 fasta -osformat2

fasta $j
                                    echo "Spliting of sequence $j completed .."
                                    arr3=`ls *.fasta`
                                    for k in $arr3
                                    do
                                            if [ $k ]
                                            then
# EXECUTING TRUST PROGRAM
echo "Executing trust program on sequence $k ..."

java -Xmx256m  nl.vu.cs.align.SelfSimilarity -fasta $k  -matrix
/home/satya/PRO_repeats/trust/Align/BLOSUM62 -noseg -o
/home/satya/PRO_repeats/trust/Align/output/ -max 38000 -gapo 15 -gapx 6 -
force >./repeats/$k
echo " "
                                            fi
                                    done
                                    mv *.fasta all_sequences
                                    cd repeats
                                    arr4=`find -size -50\c`
                                    for l in $arr4
                                    do
                                            if [ $l ]
                                            then

                                            rm $l
                                            fi
                                    done
                                    arr5=`ls *.fasta`
                                    more *.fasta >$j.all_repeats.txt
                                    cp $j.all_repeats.txt ../
```

```
cd ..
                cd all_sequences
                for m in $arr5
                do

                  if [ $m ]
                    then
                  cp  $m ../
                    fi
                done
                        cd ..
                        mv *.fasta rep_sequences
                        cd rep_sequences
                        no_files=`ls *.fasta | wc -l`
                        array=`ls *.fasta`

                   no_dirs=`expr $no_files / 490 + 1`
                   for (( x = 1; x <= $no_dirs ; x++ ))
                                do
                                mkdir seq_set$x
                        done
                                y=1
                                count=1
                                for i in $array
                                do
                            if [ $i ]
                                then
                            mv $i seq_set$y
                            count=`expr $count + 1`
                                p=`expr $count % 490`
                            if test  $p = 0
                                then
                                    y=`expr $y + 1`
                                    fi
                            fi
                                done
                        for (( y = 1; y <= $no_dirs ; y++ ))
                                do
                                cd seq_set$y
                                less *.fasta > seq_set$y.seq
                                cp seq_set$y.seq ../
                                cd ..
                                done
```

```
                                        cd ..
                                cd ..
                        fi
                        done
                        cd ..

                fi
        done
# Display Task completion message
        echo " TASK SUCCESSFULLY COMPLETED AT"
date
```

3.2.4 Information content of TRUST output files: The output of the above script neatly categorizes the files in various directories as below.

1. *all_sequences* which consists of all individual sequence files in FASTA format.

2. *rep_sequences* files which again consists of two types of files.

    a. *all_seq* (comprises of total number of repeat sequences),

    b. *seq_setx.txt* (comprises of all the repeat sequences in FASTA format where x is 1 to 5 in text file). This will facilitate the identification analysis in further steps as described in section 3.2.5.

3. *repeats* (comprises of detected repeats and repeat types in each single sequence as separate FASTA file). In each file, a. the NCBI ID of the protein, b. the number of repeat types in the protein, c. the length of each repeat type, d. the starting and the ending amino acid numbering of each repeat are indicated in separate lines. The multiple sequence alignment of each repeat type is provided, in which gaps are indicated by "-".

4. GENOME_ID.faa file consists of the total number of sequences in FASTA format in a single file and

5. GENOME_ID.faa.all_repeats (consists of all the types of repeats detected in each sequence of the whole organism in a single text file).

3.2.5 Batch submission to SMART in normal mode: The file seq_setx.txt (where x is 1 to 5), from the folder "*rep_sequences*" was taken as input and submitted to online SMART in batch mode available at the website (http://smart.embl.de/). The results obtained were saved as a complete html file in the same folder. The file was saved in Mozilla web browser for better visualization.

3.2.6 Analysis of SMART output: The output files obtained after SMART analysis were manually separated based on their presence and absence in the SMART database. The repeats that were previously identified and already present in the SMART database, were separated into a folder "*present_in_smart*" and those repeats that were absent in SMART database were separated into another folder "*not_in_smart*". The repeats that were novel according to the SMART analysis and present in the folder "*not_in_smart*" were analyzed further.

3.2.7 Local submission to PSI-BLAST program: Often, BLAST searches are useful to detect internal repeats in proteins. A region of query sequence aligns with various regions of a subject sequence in a database. This indicates that the subject sequence has several copies of the query sequence. Therefore, for further validation of repeats identified by TRUST, BLAST searches of the novel repeat sequences were carried out. To achieve this, we have downloaded NCBI NR (*release date: April 22, 2005*) and UNIPROT (*release date: April 23, 2005*) databases and installed BLAST-2.2.10 on the local Linux computers (OS: Fedora Core-2, Pentium-IV 3.00 GHz, 1 GB RAM, 80 GB hard disk). The repeats that are not present in SMART were submitted to BLAST analysis by using the commands sh align.sh followed by sh blast.sh. All the repeats types are submitted to BLAST by manually creating the text file for each repeat type. The repeats were then searched using automatic shell scripts by PSI-BLAST

program (Altschul *et al.,* 1997) for three iterations against the NCBI NR database, and WU-BLAST2 program (Chao *et al.,* 1992) against the UNIPROT database. We have also predicted the secondary structure for the repeats using PSIPRED program (Jones, 1999). The shell script for the secondary structure prediction using PSIPRED (Jones, 1999) is also included in the program used for BLAST analysis which is shown below.

sh align.sh

```
# Declare Array
declare -a arr1
declare -a arr2
declare -a arr3
declare -a arr4

path=`pwd`
cd ../../rep_sequences/

# making directory all_seq and dumping all seq
mkdir all_seq
arr2=`ls -d */ | grep "seq_set[1234][^_]"`
        for j in $arr2
        do
                if [ $j ]
                then
                cp $j/*.fasta ./all_seq/
              fi
      done
cd $path

# changing the extension of .fasta to .rep for the trust outputs
arr3=`ls *.fasta`
rename .fasta .rep *.fasta
for k in $arr3
do
if [ $k ]
        then
        cp ../../rep_sequences/all_seq/$k .
    fi
    done
```

```
# for creating folders , getting names into an array
mkdir temp
      cp *.rep ./temp
      rename .rep . ./temp/*.rep
      arr4=`ls temp`
      for l in $arr4
      do
      if [ $l ]
            then
            mkdir $l
            mv  ${l}[r.f_]*  $l
      fi
      done
      rm -rf ./temp
      cd ..
```

sh blast.sh

```
                  for j in $arr2
                  do
                  if [ $j ]
                  then
                        cd $j
# Get list of all FASTA file of child directory
                        arr3=`ls *.fasta`
# Taking fasta file
                        for k in $arr3
                        do
                        if [ $k ]
                        then
                        echo "Prediction of secondary structure for  sequence $k
is in progress ....."
                        runpsipred $k
                        rename .horiz .ss *.horiz
                        fi
                        done
                        arr4=`ls *.txt`
                        for l in $arr4
                        do
                        if [ $l ]
                        then
                        echo "PSI-BLAST of sequence $l is in progress..."
```

```
blastpgp -j 3 -h 0.001 -d /mnt/Win_E_Misc/DATABASES/ncbi_formated/nr -i
$l > $l.bla

                        echo "Wu blast of sequence $l is in      progress..."
blastall -p blastp -G 10 -E 2 –d
/mnt/Win_D_Users/Linux_partition/uniprot_databases/uniprot/uniprot_complet
e_database.fasta -i $l >$l.wu
                        fi
                        done

                        echo " "
                        cd ..
                fi
                done
echo " Task completed successfull"
```

The offline BLAST and secondary structure prediction output consist of three text files. They are PROTEIN_ID.rep.blast (PSI-BLAST output file), PROTEIN_ID.rep.wu (WU-BLAST2 output file) and PROTEIN_ID.rep.ss (PSIPRED output file). The BLAST repeats were separated manually into two folders, files that have repeats are separated into "*blast_repeats*" folder and those that have no repeats are separated into another folder "*blast_not_repeats*".

The repeats identified by BLAST i.e. the repeats in "*blast_repeats*" were submitted to offline INTERPRO and PFAM databases by using sh int_pfam.sh command.

3.2.8 Local submission to INTERPRO and PFAM databases: The INTERPRO and PFAM databases were downloaded and installed on the local Pentium IV computers. The proteins confirmed to comprise repeats according to the BLAST program were retained and searched for presence in the offline versions of INTERPRO (*Database: iprscan_DATA_10.0, Applications: iprscan_V4.1, iprscan_binn4.x_Linux*) and PFAM (*release date: April 26, 2005*) databases using sh int_pfam.sh command. The script is as follows:

sh int_pfam.sh

```
# Declare Array
declare -a arr1
declare -a arr2
declare -a arr3
# Get list of all Directory in Array 'arr1'
arr1=`ls`
# moving in child directories
for i in $arr1
do


if [ $i ]
      then
            cd $i
# Get list of all FASTA file of child directory
            arr2=`ls *.fasta`
# Taking fasta file
            for j in $arr2
            do
            if [ $j ]
           then
echo " Interpro scan for sequence $j in progress..."
iprscan -cli -i $j  -o ${j}.int -nocrc -iprlookup -format txt -verbose
echo "Blasting pfam for sequence $j in progress.."
blastall -p blastp -G 10 -E 2 -d /mnt/Win_E_Misc/DATABASES/pfam/Pfam-
A.fasta -i $j >${j}.pfm

                        fi
                       done


# Get list of all FASTA file of child directory
        arr3=`ls *.cl`
# Taking fasta file
        for k in $arr3
        for
        if [ $k ]
        then
        echo "Multiple sequence Alignment of file $k in progress..."

clustalw $k
        fi
        done
```

```
        rm -rf *.dnd
        echo " Leaving folder $j"
cd ..
fi
done
```

The output files obtained after INTERPRO and PFAM analysis were manually separated. The repeats that were already identified were separated into "*present_in_int_pfam*" folder and those that were unidentified previously were separated into "*not_in_int_pfam*" folder.

3.2.9 Online submission to INTERPRO and PFAM databases: A final check was made using online versions of INTERPRO available at the website (http://www.ebi.ac.uk/InterProScan/) and PFAM available at website (http://www.sanger.ac.uk/Software/Pfam) databases.

3.2.10 Identification and separation of novel repeats and domains: The repeats which are not present in any of these databases were considered to be novel. The novel regions comprising repeats were classified as either repeats or domains, depending upon (1) the number of times they occur in the protein sequences and (2) length of the amino acid sequence region. A repeat exists always as multiple copies in proteins and often present in tandem and comprise less than 55 amino acid residues. All the copies are required for its folding in a 3-D space. A domain can exist as a single copy in a protein and often comprises greater than 55 amino acid residues and therefore is a structurally independent folding unit.

3.2.11 Analysis of novel repeats and domains: The novel repeats and domains identified by TRUST were subjected to online PSI-BLAST analysis in order to identify other proteins from nr databases that comprise these repeats and domains. Multiple sequence alignment program, CLUSTALW (Thompson *et al.*, 1994) was used to detect the extent of sequence conservation. The

secondary structure predictions were carried out using PHD (Rost, 1994) and PSIPRED (Jones, 1999) methods.

3.2.12 Case Study: *Bacillus anthracis* str. Ames proteome: As a case study, we explain the repeat identification method by TRUST, implemented in our laboratory to identify and analyze the repeats in the *B. anthracis* str. Ames proteome.

The complete genome sequence NC_003997.faa was downloaded from NCBI website (http://www.ncbi.nlm.nih.gov/Bacteria/Bacillus_anthracis_ames) in FASTA format. The NC_003997.faa is the ID assigned by NCBI to the *B. anthracis* str. Ames proteome. The size of the file was 1.84MB and was submitted to TRUST program. The TRUST output comprised of three folders and two text files. The folders are "*all_sequences*", "*rep_sequences*" and "*repeats*". The two text files are NC_003997.faa.txt and NC_003997.faa.all_repeats.txt. The first folder, "*all_sequences*", consists of 5311 protein sequence files in FASTA format which constitute the total number of protein sequences in *B. anthracis* str. Ames proteome. The second folder, "*rep_sequences*", consists of three folders, these are "*all_seq*", "*seq_set1*" and "*seq_set2*" and two text files called "seq_set1.txt" and "seq_set2.txt". The folder, "*all_seq*", consists of 905 protein sequences that comprised of repeats. The *seq_set1* folder consists of 489 and the *seq_set2* consists of 416 repeat sequences. The text files seq_set1.txt and seq_set2.txt that comprised of protein sequences with repeats in a single text file were submitted to online SMART database. The third folder, "*repeats*" consists of 905 text files with the repeat information for each protein.

```
>NP_845711.1
# type_of_the_repeat
REPEAT_TYPE 1
# profile_length
REPEAT_LENGTH 47
# The list of repeats in the format:
# START LENGTH [PVALUE [SCORE]]
196 47      # Repeat 1
249 47      # Repeat 2
296 47      # Repeat 3
# The multiple alignment of repeats
# lo-case letters: not a part of alignment
# Profile pattern, "X": profile column, "-": a gap
# XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
>Repeat 1
GSGPRHITFHPNGKYAYVMTELSSEVIMLTYNPAEGPFTELQYISTI
>Repeat 2
NNQGSAIHISSDGRFVYAGNRGHNSIAVFSVDENSGKLTFVAHTSTE
>Repeat 3
GNWPRDFVLDPTEKFLVATNEKSHNLVLFSRSESTGELTLLQSDVAV
# type_of_the_repeat
REPEAT_TYPE 2
# profile_length
REPEAT_LENGTH 44
# The list of repeats in the format:
# START LENGTH [PVALUE [SCORE]]
45 44 # Repeat 1
104 45      # Repeat 2
# The multiple alignment of repeats
# lo-case letters: not a part of alignment
# Profile pattern, "X": profile column, "-": a gap
# XXXXXXXXXXXXXXXXXXXXXXXXXXXX-XXXXXXXXXXXXXXXXX
>Repeat 1
NPTYVTINRNNEYLYSVVKEGESGGVA-AYSINSKTGELTEENRQ
>Repeat 2
NHTVVTANYHKGTIESFVVNEEDGTVSpAASIMAHEGSGPNKERQ
# end of the protein NP_845711.1
//
```

The seq_set1.txt and seq_set2.txt from the folder "*rep_sequences*" were taken as input and submitted to online SMART analysis in batch mode. The results obtained were saved as a complete html file in Mozilla web browser as seq_set1.html and seq_set2.html files respectively. The 905 repeats were manually separated by checking them in the seq_set1.html and seq_set2.html files based on their presence and absence in the SMART database. Out of 905 repeats that were identified by TRUST, 602 repeats were already present in the SMART database and were separated into another folder "*present_in_smart*" and 303 repeats were absent in SMART database and were separated into a folder "*not_in_smart*". The repeats that were "*not_in_smart*" were edited manually by copying each repeat into another new text file in the same folder by removing the gaps in the amino acid sequence of the repeat region and saved them under the same NCBI ID of the repeat, such as np_844677.1.rep.txt. When the number of repeat types were two or more in a protein sequence, they were saved as repeat_type1, repeat_type2, repeat_type3, etc. The repeats were submitted to offline PSI-BLAST for three iterations and WU-BLAST2 analysis. The secondary structure was also predicted for the repeat using PSIPRED by using the commands sh align.sh followed by sh blast.sh.

The offline BLAST and secondary structure prediction output files consist of three text files. They are PROTEIN_ID.rep.blast (PSI-BLAST output file), PROTEIN_ID.rep.wu (WU-BLAST2 output file) and PROTEIN_ID.rep.ss (PSIPRED output file). Out of the 303 sequences submitted to BLAST, 249 sequences were also identified as BLAST repeats and we separated them manually into new folder "*blast_repeats*" and the remaining 54 into another folder "*blast_not_repeats*".

The repeats identified by BLAST were submitted to offline INTERPRO and PFAM databases by using sh int_pfam.sh command. The output files obtained after INTERPRO and PFAM analysis were manually separated. The repeats that were already known were separated into "*present_in_int_pfam*"

folder and those that were identified so far into "*not_in_int_pfam*" folder. Out of 249 BLAST repeats, we have identified 194 repeats that were present in INTERPRO and PFAM databases and 55 repeats were not identified. The INTERPRO and PFAM output files consist of two text files, PROTEIN_ID.fasta.int and PROTEIN_ID.fasta.pfam. The 55 repeats were again submitted to online INTERPRO and PFAM databases out of which we have identified 14 repeats that were novel. These were further subjected to online PSI-BLAST analysis in order to identify other proteins from nr databases that comprise these repeats and domains.

In the process, we have identified and analyzed 4 repeats and 10 domains in *B. anthracis* str. Ames proteome. We have also predicted the secondary structure for these novel repeats and domains, and function when possible. The novel repeats and domains of *B. anthracis* str. Ames identified are discussed in detail in Chapter 4.

We have also identified and analyzed 56 domains and 38 repeats in 13 archaeal organisms and 7 domains and 18 repeats of human proteome. The novel repeats and domains identified in 13 archaeal organisms are discussed in detail in Chapter 5 and human proteome in Chapter 6.

Our findings will aid in protein structure predictions by correlating the amino acid stretches with the repeats and domains identified in this project. Information obtained in this study on novel repeats and domains will be used for annotation in the databases. The tools developed in the process will also save a large amount of time and labor involved in similar studies.

3.3 Conclusions

1. TRUST is used as the main program for the novel repeat identification method.

2. We can submit up to 5 organisms simultaneously for repeat identification using TRUST as long as the total size of the file does not exceed 2MB which saves a large amount of time.

3. Analysis of TRUST output files are divided systematically for each organism which facilitates the repeat identification analysis.

4. The offline shell scripts i.e. sh blast.sh and sh int_pfam.sh comprises of more than two programs within the same script which will reduce the time and labor involved in the analysis.

5. TRUST predicts multiple repeat types with varying intervening segments within a single sequence. It showed a higher accuracy and sensitivity of repeat prediction.

3.4 References

Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. (1990). Basic local alignment search tool. *J. Mol. Biol.* **215,** 403-410.

Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucl. Acids Res.* **25,** 3389-3402.

Andrade, M. A., Ponting, C. P., Gibson, T. J. & Bork, P. (2000). Homology-based method for identification of protein repeats using statistical significance estimates. *J. Mol. Biol.* **298,** 521-537.

Chao, K. M., Pearson, W. R. & Miller, W. (1992). Aligning two sequences within a specified diagonal band. *Comput. Appl. Biosci.* **8,** 481-487.

Heger, A. & Holm, L. (2000). Rapid automatic detection and alignment of repeats in protein sequences. *Proteins,* **41,** 224–237.

Heringa, J. & Argos, P. (1993). A method to recognize distant repeats in protein sequences. *Proteins*, **17,** 391–341.

Jones, D. T. (1999). Protein secondary structure prediction based on position-specific scoring matrices. *J. Mol. Biol.* **292,** 195-202.

Mott, R. (2000). Accurate formula for P-values of gapped local sequence and profile alignments. *J. Mol. Biol.* **300,** 649-659.

Rost, B., Sander, C. & Schneider, R. (1994). PHD-an automatic mail server for protein secondary structure prediction. *CABIOS,* **10,** 53-60.

Szklarczyk, R. & Heringa, J. (2004). TRUST: Tracking Repeats Using Significance and Transitivity. *Bioinformatics,* **00,** 1-7.

Thompson, J. D., Higgins, D. G. & Gibson, T. J. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* **22,** 4673-4680.

Waterman, M. S. & Eggert, M. (1987). A new algorithm for best subsequence alignments with application to tRNA-rRNA comparisons. *J. Mol. Biol.* **197**, 723-728.

CHAPTER 4

Identification and Analysis of Novel Amino Acid Sequence Repeats and Domains in *Bacillus anthracis* str. Ames Proteome

4.1 Introduction

The anthrax is a disease of herbivores and other mammals including humans, caused by the *Bacillus anthracis* str. Ames, a Gram-positive, rod-shaped, non-motile, spore-forming bacterium (Okinaka *et al.,* 1999). It is an endospore-forming bacterium that causes inhalational anthrax. During the course of disease, endospores are taken up by alveolar macrophages where they germinate in the phagolysosomal compartment. Vegetative cells then escape from the macrophage, eventually infecting blood. Expression of the major plasmid-encoded virulence determinants, tripartite toxin and a poly-D-glutamic acid capsule are essential for full pathogenicity (Dixon *et al.,* 1999). Key virulence genes found on plasmids are pXO1 and pXO2 (Okinaka *et al.,* 1999). The 60 MDa plasmid pXO2 carries genes required for the synthesis of an antiphagocytic poly-D-glutamic acid capsule (Uchida *et al.,* 1985). The 110 MDa plasmid pXO1 (Uchida *et al.,* 1986) is required for the synthesis of the anthrax proteins, edema factor, lethal factor and protective antigen. These proteins act in binary combinations to produce two anthrax toxins: edema toxin (a protective antigen and edema factor) and lethal toxin (a protective antigen and lethal factor) (Leppla *et al.,* 1995). The chromosome encodes potential virulence factors that include haemolysins, enterotoxins, phospholipases, proteases, metalloproteases and iron-acquisition proteins.

The chromosome of *B. anthracis* str. Ames contains three homologous of sortase transpeptidase that is responsible for attachment of secreted proteins to peptidoglycan on the cell surface of Gram-positive bacteria (Pallen *et al.,* 2001). A range of important surface proteins, including enzymes and virulence related MSCRAMMs (microbial surface components recognizing adhesive matrix molecules) are anchored to the cell wall in Gram-positive bacteria by sortase, a transpeptidase in *Staphylococcus aureus* that cleaves polypeptides at a conserved LPxTG motif near the carboxyl terminus and covalently links them

to penta-glycine crossbridges in peptidoglycan (Patti *et al.,* 1994; Navarre & Schneewind, 1999). Nearly 34 candidate surface proteins which have sortase attachment sites and S-layer homology SLH domains were identified. Two putative *B. anthracis* str. Ames sortase attached genes have internalin like repeats (Guttmann & Ellar, 2000). The chromosome of *B. anthracis* str. Ames also contains the *csaAB* genes for binding of proteins with S-layer homology (SLH) domains to polysaccharide. This domain is a repetitive modular element that is present in several bacterial cell surface proteins and is involved in non-covalent association with peptidoglycan associated polymers (Lupas *et al.,* 1994). This domain comprises 55 amino acid residues (Lupas, 1996) and the potential role of most proteins with SLH domains on the surface of *B. anthracis* str. Ames is unknown (Read *et al.,* 2003). However, these surface proteins may mediate unknown interactions between *B. anthracis* str. Ames and its external environment and could be targets for vaccine and drug design. Read *et al.,* 2003, reported the complete genome sequence of *B. anthracis* str. Ames. It comprises 5,227,293 base pairs and 5,508 genes with an overall G+C content of 35.4%. Of these, 2,762 are functional genes, 1,212 are conserved hypothetical genes, 657 genes are of unknown function and 877 genes are annotated as hypothetical proteins.

The *B. anthracis* str. Ames proteome consists of several known repeats and domains. Some of these domains are as follows: 1) BRCT (Breast Cancer Carboxy terminal) domain was first identified as 100 amino acid tandem repeat at the C-terminus of the tumor suppressor gene product BRCA1, in which the germline mutations lead to nearly 50% familial breast cancer. Most BRCT domain containing proteins participate in DNA damage checkpoint or DNA repair pathways and transcription regulation (Yu *et al.,* 2003). The BRCT is an evolutionarily conserved module that exists in a large number of proteins from prokaryotes to eukaryotes. 2) Excalibur (extracellular calcium binding) domain consists of a conserved DxDxDGxxCE motif, which is strikingly similar to the

Ca^{2+} binding loop of the calmodulin like EF hand domains, suggesting an evolutionary relationship. 3) Cna_B domain forms a stalk in *S. aureus* collagen binding protein that presents the ligand binding domain away from the bacterial cell surface. 4) CBS (cystathionine beta synthase) domain is a small intracellular module with 60 amino acid residues, mostly found in two or four copies within a protein and occurs in several proteins in all kingdoms of life. Tandem pairs of CBS domains can act as binding domains for adenosine derivatives. In some cases, CBS domains may act as sensors of cellular energy status by being activated by AMP and inhibited by ATP. 5) Par B (Par B like nuclease) domain cleaves single stranded DNA, nicks supercoiled plasmid DNA and exhibits 5`-3` exonuclease activity. 6) KH (K homology) domain comprises 70 amino acid residues and is involved in RNA binding. 7) PAS and PAC domains comprise 300 and 45 amino acid residues respectively and mediate signal transduction. 8) PASTA domain is an extracellular module comprising 70 amino acids residues that folds into a globular architecture consisting of 3 β strands and a α helix which aids in penicillin binding. 9) NEAT (near transporter) domain is a 125 amino acid residues conserved region consisting mainly β strands. The NEAT domain appears to be associated with iron transport in several Gram-positive species, some of them are pathogenic.

The repeats present in *B. anthracis* str. Ames proteome are as follows: 1) RHS repeats are 21 amino acids residues long and are involved in carbohydrate binding. 2) TPR (Tetratricopeptide repeats) are 34 amino acids residues long and are involved in protein-protein interactions. 3) EZ_HEAT repeats are 37-47 amino acids residues long and occur in tandem in a number of cytoplasmic proteins that are involved in intracellular transport process. Arrays of HEAT repeats consist of 3 to 36 units forming a rod-like helical structure and appear to function as protein-protein interaction surfaces. 4) Ankyrin repeats are about 33 amino acids residues long and occur in atleast four consecutive copies; the core of the repeat appears as a helix-loop-helix structure and is involved in

protein-protein interactions. 5) LRR (lecuine rich repeats) are 20 amino acids residues long, each repeat consists of a β strand and α helix, that are oriented in an antiparallel manner. The function of LRRs includes signal transduction, transmembrane receptors, DNA repair, cell adhesion and extracellualr matrix proteins (Kobe & Deisenhofer, 1994).

As the complete genome sequence of *B. anthracis* str. Ames is available (Read *et al.,* 2003), we intended to systematically identify and analyze all the amino acid sequence repeats in this proteome. We have identified 4 repeats and 10 domains that are novel in the proteome of *B. anthracis* str. Ames. Further analysis corresponding to searches of the completed and unfinished genome databases identified some of these to be present in other bacterial genomes.

4.2 Methods

Various methods used to carry out the repeat and domain identification have been discussed in Chapter 3.

4.3 Results and Discussion

From the analysis of *B. anthracis* str. Ames proteome using TRUST program, we identified 905 proteins comprising of amino acid sequence repeats. SMART database analysis identified that 303 entries do not have a SMART description. Further based on their absence in the INTERPRO and PFAM databases and the length of repeat sequence (greater than 25 amino acid residues), we have identified about 120 proteins (data not shown) in the *B. anthracis* str. Ames proteome to comprise novel amino acid sequence repeats. We have added an additional constraint that the repeats identified by TRUST program should also be identified as a repeat by the BLAST program. Subsequent online INTERPRO and PFAM searches confirmed that these domains and repeats have not been reported before. In this work, we have identified 4 repeats and 10 domains, that are not within or part of previously reported repeats and our findings are therefore novel. Further database searches identified some of these in the proteins of other bacterial genomes. The conserved amino acid residues observed from multiple sequence alignments using the CLUSTALW program were used to describe the sequence motifs characteristic of these novel repeats and domains. Often, more than one sequence motif is associated with repeats or domains and the amino acid sequence patterns characteristic of these repeats are represented according to the PROSITE description (Falquet *et al.*, 2002).

In this work, we identified 4 repeats and 10 domains that have not been reported before in the *B. anthracis* str. Ames proteome. The repeats and domains described in 1 to 6 and 9 are also present in some other bacterial organisms, 7, 8, 10 and 11 are *Bacillus* specific, 12 and 13 are *B. anthracis* str. Ames specific. Lists of the proteins containing these novel repeats and domains are shown in Tables 4a to 4k. These tables indicate the protein identifiers (GENE or Swall_ID), the number of amino acid residues in the protein, a

description of the protein and other well characterized repeats and domains present in the protein. Some sequences representing these repeats or domains share lower than 15% pair-wise sequence identity. However, these sequences retain the conserved motifs and the positions of secondary structure elements in the multiple sequence alignment. For all the proteins, the amino acid sequence corresponding to each representative repeat and domain are shown in the multiple sequence alignments (see Figures from 4.1a to 4.1m). Conservation of the position of secondary structural elements is indicated from the multiple sequence alignment. The schematic figures used to represent these repeats and domains are shown in Figures 4.2a to 4.2m. These figures (drawn to an approximate scale) reflect the relative proximity and location of individual repeats and domains along the primary sequence. We discuss each of these novel repeats and domains below.

1. 57 amino acid residue PxV domain: The 251 amino acid residues protein corresponding to the GENE_ID BA2292 and described as hypothetical protein comprises of a 57 amino acid residues region as two copies. Further PSI-BLAST searches using sequence corresponding to the region (65-121) as a query identified 24 proteins that are described as hypothetical (see Table 4a). This region occurs as four copies in proteins from *S. amazonensis* and *H. marismortui,* as two copies in proteins from *B. anthracis, B. cereus, B. halodurans, B. thuringiensis, B. thuringiensis serovar*, *T. thermopilus, C. aurantiacus, C. aggregans, Exiguobacterium salinarium, B. weihenstephanensis, R. castenholzii, C. novyi, H. aurantiacus* and as single copy in *A. variabilis;* we therefore, describe this region as a domain. The length of proteins varied between 196 to 488 amino acid residues. The multiple sequence alignment corresponding to this domain is associated with PxV sequence motif (where x is any amino acid residue) as shown in the Figure 4.1a. The pair-wise identities between sequences corresponding to PxV domain

varied between 15-96%. The secondary structure corresponding to PxV domain is predicted to comprise 4 β strands as shown in the Figure 4.1a. The domain architecture corresponding to proteins comprising the PxV domain is shown in the Figure 4.2a.

2. 122 amino acid residue FxF domain: The 293 amino acid residues protein corresponding to the GENE_ID BA0881 and described as conserved domain protein comprises of a 122 amino acid residue region as two copies. Further PSI-BLAST searches using sequence corresponding to the region (55-176) as a query identified 10 proteins (see Table 4b). The proteins comprising this region are described as either conserved or hypothetical proteins. This region occurs as two copies in the proteins of *B. anthracis, B. cereus, B. thuringiensis, G. kaustophilus, C. tetani, C. novyi* and *D. reducens* genomes. The length of proteins varied between 262 to 305 amino acid residues. The multiple sequence alignment corresponding to this domain is associated with characteristic sequence FxF motif (Figure 4.1b) and we refer to this as the FxF domain. The pair-wise sequence identities corresponding to this domain varies between 18-97%. The secondary structure corresponding to FxF domain is predicted to comprise 1 α helix and 5 β strands and the domain architecture of proteins comprising this domain is shown in Figure 4.2b.

3. 111 amino acid residue YEFF domain: The 510 amino acid residues protein corresponding to the GENE_ID BA3695 and described as a S-layer protein comprises of a 111 amino acid residues region that is present as two copies. Further PSI-BLAST searches using sequence corresponding to the region (247-357) as a query, identified 13 proteins (see Table 4c), that are described as S-layer proteins, hypothetical or lipoproteins and correspond to the *B. anthracis* strains Ames and A2012, *B. cereus, B. thuringiensis, B. thuringiensis serovar israelensis* and *E. faecalis* genomes. The length of proteins varied between 321 to 510 amino acid residues. Five proteins

corresponding to the GENE_ID BA3695 and Bant_01004347 of *B. anthracis*, BCE_G9241_3590 and BCZK3337 of *B. cereus* and BT9727_3386 of *B. thuringiensis* comprise three copies of SLH domain, indicating a cell surface role for these proteins. This domain is characterized by conserved sequence motifs; YEFF, RGD, FTY, GKD and FVEH. We refer to this 111 amino acid region as the YEFF domain. The pair-wise sequence identities corresponding to the YEFF domain varied between 36-96%. The consensus secondary structure predicted for this domain suggests mainly β strands and the conserved sequence motifs i.e., YEFF and FTY are associated with β strands, see Figure 4.1c. The domain architecture of proteins comprising this domain is shown in the Figure 4.2c. It is intriguing that each domain comprises a RGD sequence motif, which is found in the proteins of extracellular matrix. Many viruses enter their host cells via the RGD motif-integrin interaction and synthetic peptides containing this RGD motif are active modulators of cell adhesion (Akula *et al.,* 2002). The RGD motif was originally identified as the sequence within fibronectin that mediates cell attachment. This motif has now been found in numerous other proteins and supports cell adhesion. The integrins, a family of cell surface proteins, act as receptors for cell adhesion molecules. A subset of the integrins recognize the RGD motif within their ligands, the binding of which mediates both cell substratum and cell-cell interactions (D'Souza *et al.,* 1991). The presence of RGD motif and SLH domain imply that the YEFF domain comprising proteins are also present on the cell surface and mediate protein-protein interactions.

4. 109 amino acid residue IMxxH domain: The 266 amino acid residues protein corresponding to the GENE_ID BA1021 and described as hypothetical protein comprises of a 109 amino acid residues region as two copies. Further PSI-BLAST searches using sequence corresponding to the region (4-112) as a query identified 22 proteins (see Table 4d) that are described as either

conserved or hypothetical proteins. This domain region occurs as two copies in all the proteins of *B. anthracis, B. cereus, B. thuringiensis, B. weihenstephanensis, C. acetobutylicum, C. perfringens, C. tetani, C. thermocellum, D. hafniense, C. phytofermentans,* and *A. metalliredigenes* and as single domain in the 171 amino acid residue protein BcerKBAB4DRAFT_0307. The length of proteins varied between 171 to 321 amino acid residues. The multiple sequence alignment corresponding to this domain identified the characteristic sequence motifs; IMxxH, REA and we refer to this as the IMxxH domain. The IMxxH sequence motif occurs at the N-terminal region of the domain. The pair-wise sequence identities corresponding to the IMxxH domain varied between 5-98%. The secondary structure corresponding to IMxxH domain is predicted to comprise 4 α helices as shown in Figure 4.1d. The domain architecture corresponding to proteins comprising this domain is shown in Figure 4.2d.

5. 103 amino acid residue VxxT domain: The 349 amino acid residues protein corresponding to the GENE_ID BA4716 and described as germination protein comprises of a 103 amino acid residues region as two copies. Further PSI-BLAST searches using sequence corresponding to the region (67-169) as query identified 23 proteins (see Table 4e). The proteins comprising this domain are described as germination proteins as the *B. anthracis* is an endospore forming bacterium. This domain region occurs as two copies in proteins of *B. anthracis* str. Ames, *B. cereus, B. clausii, B. thuringiensis*, *B. thuringiensis serovar israelensis, A. metalliredigene* and *B. weihenstephanensis* genomes and only once in the proteins of *S. wolfei str. Goettingen, M. thermoacetica, C. thermocellum, B. subtilis,* and *P. thermopropionicum* genomes. The length of proteins varied between 195 to 377 amino acid residues. The multiple sequence alignment corresponding to this domain identified VxxT as sequence motif. This sequence motif occurs in the N-terminal region of each protein and the pair-wise sequence identity varied between 11-98%. The secondary structure is

predicted to comprise of 2 α helices and 3 β strands as shown in Figure 4.1e. The domain architecture corresponding to proteins comprising this domain is shown in Figure 4.2e.

6. 84 amino acid residue ExW domain: The 246 amino acid residues protein corresponding to the GENE_ID BA4310 and described as hypothetical protein comprises of an 84 amino acid residues region as two copies. Further PSI-BLAST searches using sequence corresponding to the region (45-128) as a query identified 25 proteins (see Table 4f) that are described as either conserved or hypothetical proteins. This domain region occurs as two copies in proteins of *B. anthracis* str. Ames, *B. cereus, B. halodurans* (GENE_ID BH0678), *B. thuringiensis*, *B. thuringiensis serovar israelensis, G. kaustophil*us, *B. weihenstephanensis,* and *E. sibiricum* genomes and as single copy in proteins of *B. clausii*, *B. halodurans* (GENE_ID BH0983), *B. licheniformis, B. subtilis, Exiguobacterium salinarium* and *O. ihenyensis* genomes. The length of proteins varied between 142 to 273 amino acid residues. The multiple sequence alignment corresponding to this domain identified ExW as sequence motif. The pair-wise sequence identities corresponding to the ExW domain varied between 14-98%. The secondary structure of this domain is predicted to comprise 5 β strands and the conserved sequence motif is associated with one of the β strands as shown in Figure 4.1f. The domain architecture corresponding to proteins comprising this domain is shown in Figure 4.2f.

7. 104 amino acid residue NTGFIG domain: The 232 amino acid residues protein corresponding to the GENE_ID BA2665 and described as hypothetical protein comprises of a 104 amino acid residues region as two copies in tandem. Further PSI-BLAST searches using sequence corresponding to the region (16-119) as query identified 9 hypothetical proteins comprising this domain from organisms such as *B. anthracis*, *B. thuringiensis, B. weihenstephanensis* and *B. cereus.* The protein corresponding to the GENE_ID BCZK2413 of *B. cereus* is

described as group-specific protein. The list of 9 proteins comprising this domain is shown in Table 4g. The length of proteins varied between 232 to 236 amino acid residues. This domain occurs as two copies in every protein of the bacillus species as shown in the Table 4g. We refer to this as the NTGFIG domain based on the conserved sequence motif that is present at the N-terminal part. The pair-wise identities between sequences corresponding to this domain varied between 31-99%. The secondary structure corresponding to this domain is predicted to comprise 3 α helices and 2 β strands as shown in Figure 4.1g. The domain architecture corresponding to proteins comprising this domain is shown in Figure 4.2g.

8. 36 amino acid residue NxGK repeat: The 193 amino acid residues protein corresponding to GENE_ID BA3686 and described as hypothetical cytosolic protein comprises of a 36 amino acid residues region as two copies. Further PSI-BLAST searches using sequence corresponding to the region (94-129) as query identified 9 hypothetical proteins comprising this repeat region from the organisms *B. anthracis, B. thuringiensis, B. thuringiensis serovar israelensis, B. weihenstephanensis* and *B. cereus* (see Table 4h). The length of proteins varied between 189 to 193 amino acid residues, and also consist a SAP domain at the N-terminus, in addition to the novel repeat described here. A SAP domain consists of 2 α helices and is a DNA-binding motif that is involved in chromosomal organization (Aravind & Koonin, 2000). Therefore, we believe that these repeats might also participate in a similar function. The multiple sequence alignment corresponding to this repeat identified NxGK as sequence motif (Figure 4.1h). The pair-wise sequence identities between sequences corresponding to NxGK repeats varied between 36-97%. The secondary structure is predicted to comprise 1 α helix and the conserved sequence motif described above is also associated with α helix. The domain architecture corresponding to proteins comprising the NxGK repeats is shown in Figure 4.2h

9. 95 amino acid residue VYV domain: The 225 amino acid residues protein corresponding to the GENE_ID BA1701 and described as hypothetical protein comprises of a 95 amino acid residues region as two copies in tandem. Further PSI-BLAST searches using sequence corresponding to the region (31-125) as query identified BAS1577 protein of *B. anthracis,* RBTH_03882 protein of *B. thuringiensis serovar israelensis* and DSY3134 of *D. hafniense* Y51 that are described as hypothetical proteins. The length of proteins varied between 227 to 1674 amino acid residues (see Table 4i). In RBTH_03882, this region occurs as ten copies and in tandem. The multiple sequence alignment corresponding to this domain identified characteristic sequence motifs; GDxV, VYV (see Figure 4.1i). We refer to this region as VYV domain. The pair-wise identities between sequences corresponding to VYV domains varied between 29-95%. The secondary structure corresponding to VYV domain is predicted to comprise 5 β strands. The domain architecture corresponding to proteins comprising the VYV domains is shown in Figure 4.2i.

10. 75 amino acid residue KEWE domain: The 262 amino acid residues protein corresponding to the GENE_ID BA3147 and described as hypothetical protein comprises of a 75 amino acid residues region as three copies in tandem. Further PSI-BLAST searches using the sequence corresponding to the region (34-108) as query identified this domain in 6 proteins that are described as hypothetical proteins (see Table 4j). This domain exists as 2, 3 or 4 copies in these proteins. The length of proteins identified varied between 178 to 344 amino acid residues. The pair-wise identities between sequences corresponding to these regions varied between 22-69%. These domains are present in tandem and associated with SPY, MIN, LYP, KEWE and FWT conserved sequence motifs as indicated in the multiple sequence alignment (see Figure 4.1j). We refer to this region as the KEWE domain and the sequence motif occurs at the C-terminus of the domain. The secondary structure corresponding to KEWE domain is predicted to comprise 3 α helices as shown in Figure 4.1j. The

domain architecture corresponding to proteins comprising the KEWE domain is shown in Figure 4.2j.

11. 59 amino acid residue AFL domain: The 290 amino acid residues protein corresponding to the GENE_ID BA3065 and described as hypothetical protein comprises of a 59 amino acid residue region as two copies. Further PSI-BLAST searches using sequence corresponding to the region (13-71) as query identified that this region occurs twice in the proteins with GENE_ID's: BAS2851 and Bant_01003715 of *B. anthracis* strains and protein with GENE_ID: BcerKBAB4DRAFT_1832 of *B. weihenstephanensis* and once in the protein with GENE_ID: RBTH_02124 of *B. thuringiensis serovar israelensis* (see Table 4k). The lengths of the proteins varied between 145 to 297 amino acid residues and are described as hypothetical proteins. The multiple sequence alignment corresponding to this domain identified two characteristic sequence motifs; RFxI and AFL (see Figure 4.1k). We refer to this as the AFL domain. The sequence identities shared between AFL domains varied between 38-91%. The secondary structure corresponding to the AFL domain is predicted to comprise of 1 α helix and 2 β strands and the conserved sequence motif AFL is a part of the α helix. The domain architecture corresponding to protein comprising the AFL domain is shown in Figure 4.2k.

12. 53 amino acid residue RIDVK repeat: The 159 amino acid residues protein corresponding to the GENE_ID BA0482 and described as conserved domain protein comprises of a 53 amino acid region as two copies. PSI-BLAST searches identified these repeats to be specific to *B. anthracis* str. Ames, and therefore, is an orphan protein. The multiple sequence alignment corresponding to this repeat identified three characteristic sequence motifs: ITV, IGD and RIDVK (Figure 4.1l). We refer to this as the RIDVK repeat. The sequence identity shared between this RIDVK repeats in BA0482 is 45%. The secondary structure corresponding to the RIDVK repeat is predicted to comprise 3 β

strands. The domain architecture corresponding to protein comprising the RIDVK repeat is shown in Figure 4.2l.

13. a) 41 amino acid residue AGQF repeat and b) 42 amino acid residue GSAL repeat: The protein corresponding to the GENE_ID BA4081 comprises 462 amino acid residues and described as conserved domain protein contains two novel repeat types. The sequence length corresponding to repeat types are 41 and 42 amino acid residues and are present as two copies in BA4081. PSI-BLAST searches identified these repeats to be specific to this protein alone.

a) The sequence alignment corresponding to 41 amino acid residue repeat identified two characteristic sequence motifs: DLG and AGQF (Figure 4.1m-a). We refer to this as the AGQF repeat. The motif occurs at the C-terminal part of the repeat region. The sequence homology shared between this AGQF repeats is about 34%. The secondary structure corresponding to the AGQF repeat is predicted to comprise of 1 α helix. The domain architecture corresponding to protein comprising the AGQF repeat is shown in the Figure 4.2m.

b) The sequence alignment corresponding to the 42 amino acid residue tandem repeat identified three characteristic sequence motifs: GYI, GSAL and TING (Figure 4.1m-b) and is a glycine rich repeat. We refer to this as the GSAL repeat. The sequence homology shared between the GSAL repeats is 52%. The secondary structure corresponding to the GSAL repeat is predicted to comprise of 1 α helix and 1 β strand. The domain architecture corresponding to protein comprising the GSAL repeat is shown in the Figure 4.2m. This protein is associated with a 27 amino acid residue Ribosomal_S7 region that is sandwiched between the 41 amino acid residues AGQF repeat and the 42 amino acid residue GSAL repeat. These two repeats are specific to *B. anthracis* str. Ames and are therefore orphan proteins.

From the analysis of the *B. anthracis* proteome, we observed that the novel repeats and domains are present in all the strains, such as Ames, Ames ancestor, Sterne, and A2012 that have been sequenced. This indicates that these different strains of *B. anthracis* have diverged recently. We also observed that the domains PxV, FxF, YEFF, VxxT, ExW and VYV are present in proteins from several bacterial organisms. The domains NTGFIG, KEWE, AFL and the repeats NxGK are specific to bacillus. It is interesting to note that the domains VYV and AFL are present in all the *B. anthracis* species while absent in *B. cereus* genomes. The repeats RIDVK, AGQF and GSAL are specifically present only in *B. anthracis* str. Ames and are orphan proteins. This analysis explains the differences between the closely related *B. anthracis* and *B. cereus* genomes. The identification of these novel domains and repeats in subsequently sequenced genomes will add value to their annotation.

Table 4a. List of proteins containing the 57 amino acid residue PxV domain.

GENE_ID (number of residues)	Organism	Description	Number of PxV domains
BA2292 (251)	*Bacillus anthracis* str. Ames (B)	Hypothetical protein	2
BAS2138 (249)	*Bacillus anthracis* Sterne (B)	Hypothetical protein	2
BT9727_2076 (249)	*Bacillus thuringiensis* serovar konkukian str. 97-27 (B)	Hypothetical protein	2
BCZK2072 (249)	*Bacillus cereus* E33L (B)	Hypothetical protein	2
BCE2326 (249)	*Bacillus cereus* ATCC 10987 (B)	Hypothetical protein	2
BC2244 (249)	*Bacillus cereus* ATCC 14579 (B)	Hypothetical protein	2
BH1282 (222)	*Bacillus halodurans* C-125 (B)	BH1282 protein	2
BCE_G9241_2259 (249)	*Bacillus cereus* G9241 (B)	Hypothetical conserved protein	2
RBTH_03198 (251)	*Bacillus thuringiensis* serovar israelensis ATCC 35646 (B)	Hypothetical protein	2
TT_P0044 (221)	*Thermus thermopilus* HB27 (B)	Hypothetical conserved protein	2
TTHB089 (221)	*Thermus thermophilus* HB8 (B)	Hypothetical protein	2
Chlo02001630 (262)	*Chloroflexus aurantiacus* J-10-fl (B)	Hypothetical protein	2
ExigDRAFT_0608 (264)	*Exiguobacterium* sibiricum 255-15 (B)	Hypothetical protein	2
SamaDRAFT_3539 (469)	*Shewanella amazonensis* SB2B (B)	Hypothetical protein	4
rrnAC0576 (488)	*Haloarcula marismortui* ATCC 43049 (A)	Unknown	4
Ava_3757 (292)	*Anabaena variabilis* ATCC 29413 (B)	Hypothetical protein	1
BcerKBAB4DRAFT_2942 (249)	*Bacillus weihenstephanensis* KBAB4 (B)	Conserved hypothetical protein	2
B14911_22687 (254)	*Bacillus* sp. NRRL B-14911 (B)	Hypothetical protein	2
Bcer98DRAFT_2673 (249)	*Bacillus cereus* subsp. cytotoxis NVH (B)	Conserved hypothetical protein	2
RcasDRAFT_0590 (259)	*Roseiflexus castenholzii* DSM 13941 (B)	Surface protein from Gram-positive cocci, anchor region	2
RoseRSDRAFT_1732 (259)	*Roseiflexus* sp. RS-1 (B)	Surface protein from Gram positive cocci, anchor region	2
NT01CX_1619 (210)	*Clostridium novyi* NT (B)	Conserved hypothetical protein	2
HaurDRAFT_2803 (196)	*Herpetosiphon aurantiacus* ATCC 23779 (B)	Conserved hypothetical protein	2
CaggDRAFT_2922 (261)	*Chloroflexus aggregans* DSM 9485 (B)	Conserved hypothetical protein	2

Table 4b. List of proteins containing the 122 amino acid residue FxF domain.

GENE_ID (number of residues)	Organism	Description	Number of FxF domains
BA0881 (293)	*Bacillus anthracis* str. Ames (B)	Conserved domain protein	2
BCZK0785 (293)	*Bacillus cereus* E33L (B)	Hypothetical protein	2
BT9727_0783 (295)	*Bacillus thuringiensis* serovar konkukian str. 97-27 (B)	Hypothetical protein	2
BCE_G9241_0886 (293)	*Bacillus cereus* G9241 (B)	Conserved protein, putative	2
GK3171 (297)	*Geobacillus kaustophilus* HTA426 (B)	Hypothetical conserved protein	2
CTC00525 (279)	*Clostridium tetani* E88 (B)	Hypothetical protein	2
Bcer98DRAFT_3031 (293)	*Bacillus cereus* subsp. cytotoxis NVH (B)	Conserved hypothetical protein	2
B14911_04439 (305)	*Bacillus* sp. NRRL B-14911 (B)	Hypothetical protein	2
DredDRAFT_0533 (262)	*Desulfotomaculum reducens* MI-1 (B)	Hypothetical protein	2
NT01CX_1557 (276)	*Clostridium novyi* NT (B)	Conserved protein, putative	2

Table 4c. List of proteins containing the 111 amino acid residue YEFF domain.

GENE_ID (number of residues)	Organism	Description, other known domains	Number of YEFF domains
BA3695 (510)	*Bacillus anthracis* str. Ames (B)	S-layer protein, putative, SLH-domain (3)	2
BCZK3337 (492)	*Bacillus cereus* E33L (B)	S-layer protein, SLH-domain (3)	2
BT9727_3386 (510)	*Bacillus thuringiensis* serovar konkukian str. 97-27 (B)	S-layer protein, SLH-domain (3)	2
Bant_01004347 (510)	*Bacillus anthracis* str. A2012 (B)	Hypothetical protein, SLH-domain (3)	2
BCE_G9241_359 0 (492)	*Bacillus cereus* G9241 (B)	Lipoprotein, putative SLH-domain (3)	2
BA5326 (321)	*Bacillus anthracis* str. Ames (B)	Lipoprotein, putative	2
BT9727_4791 (321)	*Bacillus thuringiensis* serovar konkukian str. 97-27 (B)	Hypothetical protein	2
BC5098 (321)	*Bacillus cereus* ATCC 14579 (B)	Hypothetical protein	2
BCZK4809 (321)	*Bacillus cereus* E33L (B)	Hypothetical protein	2
RBTH_06214 (321)	*Bacillus thuringiensis* serovar israelensis ATCC 35646 (B)	Hypothetical protein	2
EF0374 (325)	*Enterococcus faecalis* V583 (B)	Lipoprotein, putative	2
EF0375 (321)	*Enterococcus faecalis* V583 (B)	Hypothetical protein	2
EF0376 (347)	*Enterococcus faecalis* V583 (B)	Hypothetical protein	2

Table 4d. List of proteins containing the 109 amino acid residue IMxxH domain.

GENE_ID (number of residues)	Organism	Description	Number of IMxxH domains
BA1021 (266)	*Bacillus anthracis* str. Ames (B)	Hypothetical protein	2
BAS0955 (283)	*Bacillus anthracis* Sterne (B)	Hypothetical protein	2
BCZK0933 (283)	*Bacillus cereus* E33L (B)	Hypothetical protein	2
BT9727_0941 (283)	*Bacillus thuringiensis* serovar konkukian str. 97-27 (B)	Hypothetical protein	2
BC1029 (283)	*Bacillus cereus* ATCC 14579 (B)	Hypothetical protein	2
RBTH_03050 (283)	*Bacillus thuringiensis* serovar israelensis ATCC 35646 (B)	Hypothetical protein	2
CAC3450 (307)	*Clostridium acetobutylicum* ATCC 824 (B)	Hypothetical protein	2
CPE0158 (303)	*Clostridium perfringens* str. 13 (B)	Hypothetical protein	2
CTC02189 (314)	*Clostridium tetani* E88 (B)	Conserved protein	2
CtheDRAFT_1311 (307)	*Clostridium thermocellum* ATCC 27405 (B)	Conserved hypothetical protein	2
DhafDRAFT_0725 (321)	*Desulfitobacterium hafniense* DCB-2 (B)	Conserved hypothetical protein	2
BCE_G9241_1042 (283)	*Bacillus cereus* G9241 (B)	Conserved protein	2
CbeiDRAFT_3331 (312)	*Clostridium beijerincki* NCIMB 8052 (B)	Conserved hypothetical protein	2
CphyDRAFT_3436 (305)	*Clostridium phytofermentans* ISDg (B)	Conserved hypothetical protein	2
ClosDRAFT_1658 (308)	*Clostridium* sp. OhILAs (B)	Conserved hypothetical protein	2
CdifQ_02001573 (254)	*Clostridium difficile* QCD-32g58 (B)	Hypothetical protein	2
BcerKBAB4DRAFT_3543 (283)	*Bacillus weihenstephanensis* KBAB4 (B)	Hypothetical protein	2
AmetDRAFT_1908 (272)	*Alkaliphilus metalliredigenes* QYMF (B)	Conserved hypothetical protein	2
CD1511 (304)	*Clostridium difficile* 630 (B)	Conserved hypothetical protein	2
CPF_0149 (303)	*Clostridium perfringens* ATCC 13124 (B)	Hypothetical protein	2
BcerKBAB4DRAFT_0307 (171)	*Bacillus weihenstephanensis* KBAB4 (B)	Conserved hypothetical protein	1
Bcer98DRAFT_1038 (303)	*Bacillus cereus* subsp. cytotoxis NVH 391-98 (B)	Conserved hypothetical protein	2

Table 4e. List of proteins containing the 103 amino acid residue VxxT domain.

GENE_ID (number of residues)	Organism	Description	Number of VxxT domains
BA4716 (349)	*Bacillus anthracis* str. Ames (B)	Germination protein gerM	2
gerM BT9727_4219 (349)	*Bacillus thuringiensis* serovar konkukianstr. 97-27 (B)	Germination protein	2
germ BCZK4235 (349)	*Bacillus cereus* E33L (B)	Germination protein	2
BCE4587 (349)	*Bacillus cereus* ATCC 10987 (B)	Germination protein gerM	2
BC4495 (349)	*Bacillus cereus* ATCC 14579 (B)	Germination protein germ	2
BSU28380 (366)	*Bacillus subtilis* subsp. subtilis str. 168 (B)	Germination protein gerM	2
BL00314 (369)	*Bacillus licheniformis* ATCC 14580 (B)	Spore germination protein GerM	2
BH3070 (365)	*Bacillus halodurans* C-125 (B)	Germination (Cortex hydrolysis) and sporulation	2
RBTH_05210 (349)	*Bacillus thuringiensis* serovar israelensis ATCC 35646 (B)	Germination protein germ	2
gerM (210)	*Bacillus subtilis* (B)	gerM	1
ABC2653 (377)	*Bacillus clausii* KSM-K16 (B)	Germination protein GerM	2
GK2667 (357)	*Geobacillus kaustophilus* HTA426 (B)	Germination (Cortex hydrolysis) and sporulation	2
OB2107 (352)	*Oceanobacillus iheyensis* HTE831 (B)	Germination (Cortex hydrolysis) and sporulation	2
SwolDRAFT_2302 (195)	*Syntrophomonas wolfei* str. Goettingen (B)	Hypothetical protein	1
MothDRAFT_0979 (200)	*Moorella thermoacetica* ATCC 39073 (B)	Similar to Spore germination protein	1
CtheDRAFT_0840 (299)	*Clostridium thermocellum* ATCC 27405 (B)	Hypothetical protein	1
gerM (349) ABF83609	*Bacillus thuringiensis* serovar kurstaki (B)	Spore germination protein	2
Bcer98DRAFT_3179 (348)	*Bacillus cereus* subsp. cytotoxis NVH 391-98 (B)	Germination protein GerM	2
BcerKBAB4DRAFT_4089 (349)	*Bacillus weihenstephanensis* KBAB4 (B)	Germination protein gerM	2
B14911_06091 (361)	*Bacillus* sp. NRRL B-14911 (B)	Spore germination protein	2
GAA01614 (295)	*Pelotomaculum thermopropionicum* SI (B)	Unnamed protein product	1
AmetDRAFT_1640 (332)	*Alkaliphilus metalliredigenes* QYMF (B)	Hypothetical protein	2
Moth_0516 (200)	*Moorella thermoacetica* ATCC 39073 (B)	Spore germination protein-like	1

Table 4f. List of proteins containing the 84 amino acid residue ExW domain.

GENE_ID (number of residues)	Organism	Description	Number of ExW domains
BA4310 (246)	*Bacillus anthracis* str. Ames (B)	Hypothetical protein	2
BT9727_3829 (246)	*Bacillus thuringiensis* serovar konkukian str. 97-27 (B)	Hypothetical protein	2
BCE4157 (246)	*Bacillus cereus* ATCC 10987 (B)	Hypothetical protein	2
BCZK3845 (246)	*Bacillus cereus* E33L (B)	Hypothetical protein	2
BC4088 (248)	*Bacillus cereus* ATCC 14579 (B)	IG hypothetical 17224	2
GK0969 (226)	*Geobacillus kaustophilus* HTA426 (B)	Hypothetical conserved protein	2
BSU30660 (145)	*Bacillus* subtilis subsp. str. 168 (B)	Hypothetical protein ytkA (PSPA8)	1
BL05305 (147)	*Bacillus licheniformis* ATCC 14580 (B)	Conserved protein YtkA	1
BH0983 (157)	*Bacillus halodurans* C-125 (B)	BH0983 protein	1
Bant_01004966 (252)	*Bacillus anthracis* str. A2012 (B)	Protein chain release factor A	2
RBTH_02670 (248)	*Bacillus thuringiensis* serovar israelensis ATCC 35646 (B)	Hypothetical protein	2
BCE_G9241_4093 (246)	*Bacillus cereus* G9241 (B)	IG hypothetical protein	2
OB2488 (166)	*Oceanobacillus ihenyensis* HTE831 (B)	Hypothetical conserved protein	1
ABC0230 (158)	*Bacillus clausii* KSM-K16 (B)	Unknown conserved protein	1
BH0678 (246)	*Bacillus halodurans* C-125 (B)	BH0678 protein	2
ABC4088 (142)	*Bacillus clausii* KSM-K16(B)	Hypothetical protein	1
ExigDRAFT_1796 (161)	*Exiguobacterium sibiricum* 255-15 (B)	Hypothetical protein	1
OB3282 (155)	*Oceanobacillus ihenyensis* HTE831 (B)	Hypothetical conserved protein	1
BcerKBAB4DRAFT_2040 (241)	*Bacillus weihenstephanensis* KBAB4 (B)	Conserved hypothetical protein	2
B14911_09907 (144)	*Bacillus* sp. NRRL B-14911 (B)	Hypothetical protein	1
B14911_05359 (273)	*Bacillus* sp. NRRL B-14911 (B)	Hypothetical protein	2
BAA83944 (267)	*Bacillus halodurans* (B)	Unnamed protein product	2
BH1853 (158)	*Bacillus halodurans* C-125 (B)	Hypothetical protein	1
Bcer98DRAFT_3614 (177)	*Bacillus cereus* subsp. cytotoxis NVH 391-98 (B)	IG hypothetical protein	2
ExigDRAFT_0574 (253)	*Exiguobacterium sibiricum* 255-15 (B)	Hypothetical protein	2

Table 4g. List of proteins containing the 104 amino acid residue NTGFIG domain.

GENE_ID (number of residues)	Organism	Description	Number of NTGFIG domains
BA2665 (232)	*Bacillus anthracis* str. Ames (B)	Hypothetical protein	2 (tandem)
BT9727_2444 (232)	*Bacillus thuringiensis* serovar konkukian str. 97-27 (B)	Hypothetical protein	2 (tandem)
BCZK2413 (232)	*Bacillus cereus* E33L (B)	Group-specific protein	2 (tandem)
BCE2700 (234)	*Bacillus cereus* ATCC 10987 (B)	Hypothetical protein	2 (tandem)
BC2674 (234)	*Bacillus cereus* ATCC 14579 (B)	Hypothetical protein	2 (tandem)
Bant_01003317 (236)	*Bacillus anthracis* str. A2012 (B)	Hypothetical protein	2 (tandem)
BCE_G9241_CNI_0263 (234)	*Bacillus cereus* G9241 (B)	Conserved hypothetical protein	2 (tandem)
BcerKBAB4DRAFT_0535 (232)	*Bacillus weihenstephanensis* KBAB4 (B)	Conserved hypothetical protein	2 (tandem)
Bcer98DRAFT_0128 (234)	*Bacillus cereus* subsp. cytotoxis NVH 391-98 (B)	Conserved hypothetical protein	2 (tandem)

Table 4h. List of proteins containing the 36 amino acid residue NxGK repeat.

GENE_ID (number of residues)	Organism	Description, other known domains	Number of NxGK repeats
BA3686 (193)	*Bacillus anthracis* str. Ames (B)	Hypothetical protein, SAP domain (1)	2
BT9727_3378 (193)	*Bacillus thuringiensis* serovar konkukian str. 97-27 (B)	Hypothetical protein, SAP domain (1)	2
BCZK3328 (193)	*Bacillus cereus* E33L (B)	Hypothetical protein, SAP domain (1)	2
BC3626 (193)	*Bacillus cereus* ATCC 14579 (B)	Hypothetical protein, SAP domain (1)	2
BCE3645 (193)	*Bacillus cereus* ATCC 10987 (B)	Hypothetical protein, SAP domain (1)	2
RBTH_03615 (193)	*Bacillus thuringiensis* serovar israelensis ATCC 35646 (B)	Hypothetical cytosolic protein, SAP domain (1)	2
BCE_G9241_3579 (193)	*Bacillus cereus* G9241 (B)	Hypothetical cytosolic protein SAP domain (1)	2
BcerKBAB4DRAFT_0944 (193)	*Bacillus weihenstephanensis* KBAB4 (B)	Conserved hypothetical protein SAP domain (1)	2
B14911_25780 (189)	*Bacillus* sp. NRRL B-14911 (B)	Hypothetical protein SAP domain (1)	2

Table 4i. List of proteins containing the 95 amino acid residue VYV domain.

GENE_ID (number of residues)	Organism	Description	Number of VYV domains
BA1701 (225)	*Bacillus anthracis* str. Ames (B)	Hypothetical protein	2 (tandem)
BAS1577 (227)	*Bacillus anthracis* str. Sterne (B)	Hypothetical protein	2 (tandem)
RBTH_03882 (1004)	*Bacillus thuringiensis* serovar israelensis ATCC 35646 (B)	Hypothetical exported protein	10 (tandem)
DSY3134 (1674)	*Desulfitobacterium hafniense* Y51 (B)	Hypothetical protein	2 (tandem)

Table 4j. List of proteins containing the 75 amino acid residue KEWE domain.

GENE_ID (number of residues)	Organism	Description	Number of KEWE domains
BA3147 (262)	*Bacillus anthracis* str. Ames (B)	Hypothetical protein	3 (tandem)
BAS2924 (344)	*Bacillus anthracis* str. Sterne (B)	Hypothetical protein	4 (tandem)
RBTH_06405 (331)	*Bacillus thuringiensis* serovar israelensis ATCC 35646 (B)	Hypothetical protein	4 (tandem)
pE33L466_0092 (328)	*Bacillus cereus* E33L (B)	Hypothetical protein	4 (tandem)
Bant_01003795 (178)	*Bacillus anthracis* str. A2012 (B)	Hypothetical protein	2 (tandem)
pBMB165 (247)	*Bacillus thuringiensis* serovar tenebrionis (B)	Hypothetical protein	3 (tandem)

Table 4k. List of proteins containing the 59 amino acid residue AFL domain.

GENE_ID (number of residues)	Organism	Description	Number of AFL domains
BA3065 (290)	*Bacillus anthracis* str. Ames (B)	Hypothetical protein	2
BAS2851 (297)	*Bacillus anthracis* str. Sterne (B)	Hypothetical protein	2
Bant_01003715 (293)	*Bacillus anthracis* str. A2012 (B)	Hypothetical protein	2
RBTH_02124 (145)	*Bacillus thuringiensis* serovar israelensis ATCC 35646 (B)	Hypothetical protein	1
BcerKBAB4DRAFT_1832 (291)	*Bacillus weihenstephanensis* KBAB4 (B)	Conserved hypothetical protein	2

The proteins are represented by their corresponding GENE_ID along with the number of amino acid residues indicated in brackets in the first column. The organism and corresponding phylogeny are indicated in the second column; 'A' represents archaea and 'B' represents bacteria. The third column contains the description of the proteins containing the repeats or the domains identified elsewhere, including those identified in the present work and the total number of such repeats or domains. The fourth column represents exclusively the total number of novel repeats or domains identified in this work.

Figure 4.1a: Multiple sequence alignment of 57 amino acid residue PxV domain.

```
Secondary structure                    EEEEE         EE
RcasDRAFT_0590_1(32-89)                VRVIHAS-PDAPAVDVIVNGNR--ALTNVPFFAASAYLDLPAGS
RoseRSDRAFT_1732_1(32-89)              VRVVHAS-PDAPAVDVIVNGNK--ALTNVPFFAASAYLDLPAGS
Chlo02001630_1(32-90)                  VRVIHAS-PDAPAVDVFVNGNA--VLTNVGFFAASPYLDLPAGT
CaggDRAFT_2922_1(31-89)                VRVIHAS-PDAPAVDVFVNGNA--VLTNVGFFAASPYLDLPAGT
HaurDRAFT_2803_1(4-62)                 VRVMHAS-PDAPAVDIFVDGKA--VLTSVPFFALSGQLALPDGT
B14911_22687_1(67-124)                 VRVVHAS-PDAPNVDIYVNGNR--ILKDFPYKDVSGYLSLPAGK
HaurDRAFT_2803_2(105-162)              VRVIHGS-PDAPAVDIKIAGTQN-VVVKGAKFGDAATLEVPAGT
rrnAC0576_1(67-124)                    VRVAHMS-PNAPNVDVYLEGDA--VLEDVPFGAVSQYLDVPAGE
rrnAC0576_2(284-341)                   VRVAHMS-PNAPNVDVYVDGSA--VLEDVPFGAVSDYLEVPAGA
BH1282_1(30-89)                        VRVLHAS-PDAPPVDVYIDGKK--QMEGVPFKQTSSYFNVPAGD
ExigDRAFT_0608_1(29-86)                VRVIHAS-PDAPAVDIAVDGKK--AVSGAEFKAVTDYLTLPAGE
RBTH_03198_2(161-218)                  IRFAHFS-PDTPVVNVDLKDGDH-LFENVLFKQITDFLQVSPGT
BC2244_2(159-216)                      IRFAHFS-PDTPVVNVDLKDGDH-LFENVLFKQITDFLQVSPGT
BcerKBAB4DRAFT_2942_2(159-216)         IRFAHFS-PDTPVVNVNLKDGDH-LFENVLFKQITDFLQVSPGT
BCE_G9241_2259_2(159-217)              IRFAHFS-PDTPVVNVSLKGGDH-LFENVLFKQITDFLEVSPGT
BCE_2326_2(159-216)                    IRFAHFS-PDTPVVNVSLKGGDH-LFENVLFKQITDFLEVSPGT
BA2292_2(161-218)                      IRFAHFS-PDTPVVNVSLKDGDH-LFENVLFKQITDFLEVSPGT
BAS2138_2(159-216)                     IRFAHFS-PDTPVVNVSLKDGDH-LFENVLFKQITDFLEVSPGT
BT9727_2076_2(159-216)                 IRFAHFS-PDTPVVNVSLKDGDH-LFENVLFKQITDFLEVSPGT
BCZK2072_2(159-216)                    IRFAHFS-PDTPVINVSLKDGDH-LFENVLFKQITDFLEVSPGT
Bcer98DRAFT_2673_2(159-216)            IRFAHFS-PDTSVVNVSLKNGDH-LFENVLFKQVTDYLQVSPGT
NT01CX_1619_2(113-170)                 VKFVHLS-PGTPNVDITLPNGTI-LFKDVEFEEGTDYIPLKVGT
B14911_22687_2(164-221)                ARFIHLS-PDAPAVDIAVKKGDV-IFPNISFRQATQYLGLTPMT
RcasDRAFT_0590_2(131-188)              VRVIHFS-PDAPAVDIKVAGGPT-LISNLAFPNASNYLPVDAGS
RoseRSDRAFT_1732_2(131-188)            VRVIHFS-PDAPAVDIKVAGGPT-LISNLAFPQASNYLPVDAGS
Chlo02001630_2(131-188)                VRVYHFS-PDAPAVDVKLANGTT-LISNLAFPNASDYLEVPAGT
CaggDRAFT_2922_2(130-187)              VRVYHFS-PDAPAVDVKLANGTT-LISNLAFPDASDYLEVPAGT
ExigDRAFT_0608_2(126-183)              VRVAHFA-PDAPAVDVAPKGGDP-LFSDLEFSKVSDYGTLDAGT
TTHB089_2(124-181)                     IRVVHAS-PDAPAVDVAVKGGPV-LFAGLPFPRASAYASVPAGT
TTP0044_2(124-181)                     IRVVHAS-PDAPAVDVAVKGGPV-LLAGLPFPRASAYASVPAGT
BH1282_2(130-187)                      LRAVHLS-PDTPAVQLHLSAANV-DMPSLSFENASRYIDLPAGA
RBTH_03198_1(65-121)                   IRIFHAD-PNIPAVDILVNGQKV--IKNISFKQFSPYLSLVQGK
BC2244_1(63-119)                       IRIFHAD-PNIPAVDILVNGQKV--IKNISFKQFSPYLSLVQGK
BCE_G9241_2259_1(63-119)               IRFFHSA-SNTPAVDILVNGQKV--IKNISFKQFSPYLTLVQGK
BAS2138_1(63-119)                      IRFFHSA-SNTPAVDILVNGQKV--IKNISFKQFSPYLTLVQGK
BCE_2326_1(63-119)                     IRFFHSA-SNTPAVDILVNGQKV--IKNISFKQFSPYLTLVQGK
BA2292_1(65-121)                       IRFFHSA-SNTPAVDILVNGQKV--IKNISFKQFSPYLTLVQGK
BCZK2072_1(63-119)                     IRFFHSA-SNTPAVDILVNGQKV--IKNISFKQFSPYLTLVQGK
BT9727_2076_1(63-119)                  IRFFHSA-SNTPAVDILVNGQKV--IKNISFKQFSPYLTLVQGK
BcerKBAB4DRAFT_2942_1(63-119)          MRIFHTA-PHTPAVDIIINGQKV--IKNISFKQFSPYLSLVQGK
Bcer98DRAFT_2673_1(63-119)             MRIFHAS-PHTAPVDILINGQKV--IKNITFQQFSPYFSLMQGQ
SamaDRAFT_3539(264-321)                IRVAHSA-ADVPQVDILANGTKVDALSGAAFGQASGYLNLAPGE
Ava_3757(63-120)                       LRVINAAVPTASPVDVIVNGQRV--LENVNFRQASRYVNVTPGN
TTHB089_1(24-81)                       VRVAHLS-PDAPAVDVLVNGQRA--ITGLAFKEVTPYIPLPAAK
TTP0044_1(24-81)                       VRVAHLS-PDAPAVDVLVNGQRA--ITGLAFKEVTPYIPLPAAK
NT01CX_1619_1(15-72)                   MRLLNAS-PNAPAVDVYFNGQLI--TSNLAYKEFTEYMSTSPGL
                                        :  :   .  :::              :          .
consensus/80%                          lRhhHhu.PssPsVsl.lpstt...hpsl.F.phosalpls.Gp
```

Contd…..

```
Secondary structure                  EEEEEE         EE
RcasDRAFT_0590_1(32-89)             YDIQVVPAGAT-S-PVVID 58
RoseRSDRAFT_1732_1(32-89)           YDIQVVPAGAT-S-PVVID 58
Chlo02001630_1(32-90)               YRVQVAPTGAG-AGSAVID 59
CaggDRAFT_2922_1(31-89)             YRVQVAPTGAG-AGSAVID 59
HaurDRAFT_2803_1(4-62)              YTIDIAPAGAG-VAASVFE 59
B14911_22687_1(67-124)              YQIDIYPAGDM-V-STVLS 58
HaurDRAFT_2803_2(105-162)           YSFDISPAGSS-T--VLFT 58
rrnAC0576_1(67-124)                 RSVEITAAGD--PDTSVFS 58
rrnAC0576_2(284-341)                RTVEITAAGD--PDTSVFE 58
BH1282_1(30-89)                     HMITIFAAGDDPAETPVIE 60
ExigDRAFT_0608_1(29-86)             HKVEVFAAGT--TKDPVLS 58
RBTH_03198_2(161-218)               ADIEISLANNK---NVLLT 58
BC2244_2(159-216)                   ADIEISLADNK---NVLLT 58
BcerKBAB4DRAFT_2942_2(159-216)      ADIEVSLADTK---KVLLT 58
BCE_G9241_2259_2(159-217)           ADIEVSLADNQ---NVLLT 58
BCE_2326_2(159-216)                 ADIEVSLADHQ---SVLLT 58
BA2292_2(161-218)                   ADIEVSLADNQ---SVLLT 58
BAS2138_2(159-216)                  ADIEVSLADNQ---SVLLT 58
BT9727_2076_2(159-216)              ADIEVSLADNQ---SVLLT 58
BCZK2072_2(159-216)                 ADIEVSLADNQ---SILLT 58
Bcer98DRAFT_2673_2(159-216)         ADIEISLADTK---KNLVT 58
NT01CX_1619_2(113-170)              YTIEAKPTGSD---KTVLT 58
B14911_22687_2(164-221)             VDLEVRVAGSS---NTVLS 58
RcasDRAFT_0590_2(131-188)           YDLQVTPAGGT---AVVLD 58
RoseRSDRAFT_1732_2(131-188)         YDLQVTPAGGT---AVVLD 58
Chlo02001630_2(131-188)             YDLQVTPAGGS---AVVIN 58
CaggDRAFT_2922_2(130-187)           YDLQVTPAGGD---AVVIN 58
ExigDRAFT_0608_2(126-183)           YDLEVRPAGAT---DVVKA 58
TTHB089_2(124-181)                  YDLEVRAAGTA---TVALD 58
TTP0044_2(124-181)                  YDLEVRAAGTA---TVALD 58
BH1282_2(130-187)                   YDLDIRMIETD---DVATE 58
RBTH_03198_1(65-121)                YRIDIVPVGNET---PIFS 57
BC2244_1(63-119)                    YRIDIVPVGNET---PIFS 57
BCE_G9241_2259_1(63-119)            YRIDIVPVGNET---PIFS 57
BAS2138_1(63-119)                   YRIDIVPVGNET---PIFS 57
BCE_2326_1(63-119)                  YRIDIVPVGNET---PIFS 57
BA2292_1(65-121)                    YRIDIVPVGNET---PIFS 57
BCZK2072_1(63-119)                  YRIDIVPVGNET---PIFS 57
BT9727_2076_1(63-119)               YRIDIVPVGNET---PIFS 57
BcerKBAB4DRAFT_2942_1(63-119)       YRIDIVPVGNET---PIFS 57
Bcer98DRAFT_2673_1(63-119)          YRLDIVPLDNET---PIFS 57
SamaDRAFT_3539(264-321)             YQVDTVLTSDNS---VVGI 59
Ava_3757(63-120)                    IQVLFVTSGTNS---TIAS 58
TTHB089_1(24-81)                    VRVQVVPAGQDAP--VVID 58
TTP0044_1(24-81)                    VRVQVVPAGQDAP--VVID 58
NT01CX_1619_1(15-72)                YNVKVFPHGKLSS--PIID 58

consensus/80%                       hplpl..ssst....slhs
```

BA2292 is homologous to protein GBAA2292 from *B. anthracis* str. "Ames Ancestor." BAS2138 is homologous to proteins BT9727_2076 from *B. thuringiensis* serovar konkukian str. 97-27 and Bant_01002917 from *B. anthracis* str. A2012.

Figure 4.1b: Multiple sequence alignment of 122 amino acid residue FxF domain.

```
Secondary structure                HHHHHHH         EEEEE          EEEEEE
BA0881_1(55-176)                   IYQFLHKELPRLEEYQISLSGIEIEKRDNG-YDVAVFIRSTVPKPIS
BCZK0785_1(55-176)                 IYQFLHKELPRLEEYQISLSGIEIEERDNG-YDVAVFIRSTVPKPIS
BCE_G9241_0886_1(55-176)           IYQFLHKELPRLEEYQISLSGIEIEKRDSG-YDVAVFIRSTVPKPIS
Bcer98DRAFT_3031_1(55-176)         IYQFLHKELPRLQENQISLSGIEIEKREGS-YAVAAFIRSSISKPIS
BT9727_0783_1(58-179)              IYQFLHKSLPTLQENQISLAGIESKKHENA-YYITTFIRSSVKHPIQ
GK3171_1(46-167)                   VYRFYHEQLPPLQPNQISISGVKLVEYNDG-FVAVAILRNTLPKPVR
B14911_04439_1(59-182)             VLRFLNNELPPLLPNQISLAGIELQQDGGS-VTVAAFVRSSLSKAVE
BA0881_2(185-293)                  ALRNFVDNLTPPNDGEINFLGLQAARKENGDLHTTLLIRNGCKDNIQ
BCZK0785_2(185-293)                ALRNFVDNLTPPNDGEINFLGLQAARKENGDLHTTLLIRNGCKDNIQ
BCE_G9241_0886_2(185-293)          ALRNFVDNLTPPNDGEINFLGLQAARKENGDLHTTLLIRNGCKENIQ
Bcer98DRAFT_3031_2(185-293)        ALRNFVESLTPPQNGELNFLGLQAAQKENGDLHATILIRNGCKRNIQ
BT9727_0783_2(188-295)             KLQEIIANLDPPEEDEINFRGLNAVVEENGDLNATILIRNGYNKNIT
B14911_04439_2(191-305)            KLKQMVEQMDPPKIGEINFMGIQAKVADNEDLQVTLLIRNGNDQNVM
GK3171_2(176-297)                  QLQALVDSVPPPAPGEVNFMGIEAKQLPSGELGVTLLIRNGSDKHIH
NT01CX_1557_2(164-276)             QYEKFLKELPLLREGQVTMNAYDVYTNEDDGIAVELVIRNGRHNGVD
DredDRAFT_0533_2(156-262)          QFTTFLKKLPSVQEGSINIDTYSIEKNNDGSLTVAIVLRHRLAKPTV
CTC00525_2(170-279)                VFKEFLESLPKLERGQGSISVFTITQYENGDLLMTLLVRNATDEAVT
CTC00525_1(36-159)                 LEEELREVIPKVEEGKINIAGIYAFDQGDK-VEVKAYLANGLSQKIN
NT01CX_1557_1(31-154)              CLEEELEALPAIKEGELDVN-VDFFFDLGDRYEASIFIRNGLSTGVN
DredDRAFT_0533_1(25-147)           LMQEEINNLPQITDGTVAIDSIYTVNWEDK-IEIGFYLRNVTSHKIC
                                           :         .         .         :
consensus/80%                      hhp.hhcpLs..ppsplsh.ulph.ptpss.htsshhlRsshtcslp


Secondary structure                  EEEEE      EEEEE    EEEEE
BA0881_1(55-176)                   FEEVTLILLNKEKKLCARKTFNLSALGDIPSNVNMPFIFTFEQET
BCZK0785_1(55-176)                 FEEVTLILLNKEKKLCARKTFNLSALGDIPANVNMPFIFTFEQET
BCE_G9241_0886_1(55-176)           FEEVTLILLNKEKKLCARKTFNLSALGDIPANVNMPFIFTFEQET
Bcer98DRAFT_3031_1(55-176)         FEEVTLLLLNKEDELCARKTFNLSDIGDIPANVNMPWVFTFDEET
BT9727_0783_1(58-179)              FETLTLSLLNKNGETCARQTFDLSHLEGIPSNVNMPWTFVFEENS
GK3171_1(46-167)                   FERIRLLLLDEDGTAIARKEFDMSPFGELPPMTARPWRFLFAAED
B14911_04439_1(59-182)             FKKTHLLLVGPDEEILARKEFDLTEIGEIPAKSSRPWNFTFNSSD
BA0881_2(185-293)                  LEQLPLHIEDATGAVVVKGAFTLPNLEIKAN-TTKPWSFVFPASS
BCZK0785_2(185-293)                LEQLPLHIEDATGAVVVKGAFTLPNLEIKAN-TTKPWSFVFPASS
BCE_G9241_0886_2(185-293)          LEQLPLHIEDATGAVVVKGAFTLPNLEIKAN-TTKPWSFVFPASS
Bcer98DRAFT_3031_2(185-293)        LKQLPLHIEDASGEIVVKGAFTLPNLEIKAN-STKPWSFIFPVSF
BT9727_0783_2(188-295)             LEQLPLHISDRSESTVAERIFVLKDFQIKAN-STKPWTFTFPADS
B14911_04439_2(191-305)            LQQLPLQVEDATSEVIAKGGFQLDKFELKAN-TSKPWTFIFPKSL
GK3171_2(176-297)                  FEQIPLEVRDYAGDIVARGLFPCH-LEVKAH-TSKPWTFLFPPEL
NT01CX_1557_2(164-276)             IKRIPLSIYDKDKKLVASGTFYLEDASLNPI-SAKVYLFTFSKDE
DredDRAFT_0533_2(156-262)          LSRFQFGIVDTNKSIVARAAFVIEQYILEPG-MFLLRSFKFTPET
CTC00525_2(170-279)                MTKMPITLKTQKGETILSGVFDIENFTVNPY-KARVLSLIFKKEV
CTC00525_1(36-159)                 FEDVPIYIINSKEEKLAYQVFDLSEEGDIPSGKAIPVKLNFNKQN
NT01CX_1557_1(31-154)              LEKIPFIVLDKDEKEVGRKIFNLREVGEIPARSVRPWKIYFEKDE
DredDRAFT_0533_1(25-147)           FTQTPLKILNPKGEVLASVTINLSDMGDIPAYSVRPWRFYLGKED
                                   :    : :           :        .        : :  .
consensus/80%                      hcplsLhl.stptphhscthFsLptht.hss.sshPa.FhF.tpp
```

Contd…..

```
Secondary structure
BA0881_1(55-176)                I-TDAALSQTDWELAFELESK--HTLDLDPSWEA 122
BCZK0785_1(55-176)              I-TDAALSQTDWELAFELESK--HALDLDPSWEA 122
BCE_G9241_0886_1(55-176)        I-TDADLSQTDWELAFELESK--HVLDLDPSWEA 122
Bcer98DRAFT_3031_1(55-176)      I-TDAELSQTDWQLAFELEGE--HRLDLDPTWET 122
BT9727_0783_1(58-179)           I-TEATLSNEDWQLVFELQGK--HSLDLDPIWQE 122
GK3171_1(46-167)                K-LVDQLPADGWKIAFELTPR--HRLDLEESWEQ 122
B14911_04439_1(59-182)          L-LTDSIPAEGWKLAFEIRNNEEHRLDLDEAWEN 124
BA0881_2(185-293)               I-LKEDMDLSSWKALVPQD--------------- 109
BCZK0785_2(185-293)             I-LKEDMDLSSWKALVPQD--------------- 109
BCE_G9241_0886_2(185-293)       I-LKEDMELSSWKALVPQD--------------- 109
Bcer98DRAFT_3031_2(185-293)     V-LKKEMDLSTWKAIVPQD--------------- 109
BT9727_0783_2(188-295)          V-SKEPIDLSKWKAFIPQ---------------- 108
B14911_04439_2(191-305)         L-LKDNPDLSSWKAYPLQQQVQTEI--------- 115
GK3171_2(176-297)               L-HKAEPDWTSWKVTIPSSPAQSEKQETPSSDE- 122
NT01CX_1557_2(164-276)          L-LREDYNLKNWTIQFLLNSNVN----------- 113
DredDRAFT_0533_2(156-262)       I-VNSDADINQCSIAFL----------------- 107
CTC00525_2(170-279)             VNIEEDFDLSTCKIIFERE--------------- 110
CTC00525_1(36-159)              I-LVDKIPQDDWKVVFGGNDVKGVRYVNIELESI 124
NT01CX_1557_1(31-154)           L-NVEGINLKDLKIVFDSRIKAAGVVNVQYENLP 124
DredDRAFT_0533_1(25-147)        L--TLDNSLKDLKIAFNSRNIPPYMLVIEDRLPE 123

consensus/80%                   l.hptphs.psWchhh..p..............
```

BA0881 is homologous to proteins GBAA0881 *B. anthracis* str. "Ames Ancestor," BAS0837 from *B. anthracis* str. Sterne and Bant_ 01001534 from *B. anthracis* str. A2012.

The multiple sequence alignments corresponding to representative repeats and domains from various proteins along with their GENE or SWall identifiers. (a) PxV domain, (b) FxF domain, (c) YEFF domain, (d) IMxxH domain, (e) VxxT domain, (f) ExW domain, (g) NTGFIG domain, (h) NxGK repeat (i) VYV domain, (j) KEWE domain, (k) AFL domain, (l) RIDVK repeat, (m) (a) AGQF repeat and (b)GSAL repeat. The numbers given in brackets indicate the start and end of amino acid residue positions corresponding to either the repeat or domain. The 80% consensus is labeled according to the alignment to the alignment generated at the website www.bork.embl-heidelberg.de/Alignment/consensus.html: alcohol (o, ST); aliphatic (I, ILV); any (., ACDEFGHIKLMNPQRSTVWY); aromatic (a, FHWY); charged (c, DEHKR); hydrophobic (h, ACFGHIKLMRTVWY); negative (-, DE); polar (p, CDEHKNQRST); positive (+, HKR); small (s, ACDGNPSTV); tiny (u, AGS); turn-like (t, ACDEGHKNQRST). A capital letter indicates 80% conservation of corresponding amino acid residue. The secondary structure prediction indicated at the top was derived using the PHD program. Residues predicted with greater than 82% accuracy to form β-sheets are represented by 'E' and α-helices are represented by 'H'.

Figure 4.1c: Multiple sequence alignment of 111 amino acid residue YEFF domain.

```
Secondary structure                    EEEE                    EEEE     EEEE
EF0374(62-172)                 ILSS--TDWQGTKVYDKNNNNLTAENANFIGLAKYDGETGFYEFFDKETGET
EF0375(58-168)                 ILSG--TDWQGTRVYDAAGNDLTAENANFIGLAKYDGETGFYEFFDKNTGET
EF0376(59-172)                 GLSE--KDWAGTRVYDRNGNDLTDENQNLLHAIKFDATTSFYEFFDKETGES
BA5326(58-168)                 ILSD--TNWQGTRVYDKDKNDVTKENANFIGLAKYDAKSGRYEFFDAKTGAS
BCZK4809(58-168)               ILSD--TNWQGTRVLDKDKNDLTKENANFIGLAKYDAKSGRYEFFDAKTGAS
BT9727_4791(58-168)            ILSD--TNWQGTRVYDKDKNDVTKENANFIGLAKYDAKSGRYEFFDAKTGAS
BC5098(58-168)                 ILSE--TNWQGTRVYDKDKNDLTKENANFIGLAKYDAKSGRYEFFDAKTGAS
RBTH_06214(58-168)             ILSK--TNWQGTRVYDKDKNDLTKENANFIGLAKYDAKSGRYEFFDAKTGAS
BA3695(247-357)                ILGE--TNWQGTKVYDKDHNDVTKENQNFIGLAKYDAKTARYEFFNASTGES
Bant_01004347(247-357)         ILGE--TNWQGTKVYDKDHNDVTKENQNFIGLAKYDAKTARYEFFNASTGES
BT9727_3386(247-357)           ILGE--TNWQGTKVYDKDHNDVTKENQNFIGLAKYDAKMARYEFFNASTGES
BCZK3337(229-339)              ILGE--TNWQGTKVYDKDHNDVTKENQNFIGLAKYDAKTARYEFFNASTGES
BCE_G9241_3590(229-339)        ILGE--TNWQGTKVYDKDHNDVTKENQNFIGLAKYDAKTARYEFFNAKTGES
EF0376(223-336)                FDGTPQLLWNGTKVVDKDGNDVTSANQNFISLAKFDQDSSKYEFFNLQTGET
EF0375(199-310)                ILGT--TLWNGTKVVDKNGNDVTAANQNFISLAKFDPNTSKYEFFNLQTGET
EF0374(203-314)                ILGA--TLWNGTKVLDEDGNDVTEANKMFISLAKFDNKTSKYEFFDLETGKT
BA3695(388-499)                ILSS--TLWNGTVVLDEQGNNVTKYNSNLISLAKYDKNTNKYEFFNVNTGES
BT9727_3386(388-499)           ILSS--TLWNGTVVLDEQGNNVTKYNSNLISLAKYDKNTNKYEFFNVNTGES
Bant_01004347(388-499)         ILSS--TLWNGTVVLDEQGNNVTKYNSNLISLAKYDKNTNKYEFFNVNTGES
BCZK3337(370-481)              ILSS--TLWNGTVVLDEQGNNVTKYNSNLISLAKYDENTNKYEFFNVNTGES
BCE_G9241_3590(370-481)        ILSS--TLWNGTVVLDDQGNDVTKYNSNLISLAKYDKNTNKYEFFNVNTGES
BA5326(199-310)                ILGG--TLWHGTKVLDEAGNDVTQFNSNFISLAKFDDKSNKYEFFNSETGQS
BCZK4809(199-310)              ILGG--TLWHGTKVLDEAGNDVTQFNSNFISLAKFDDKFNKYEFFNSETGQS
BT9727_4791(199-310)           ILGG--TLWHGTKVLDETGNDVTQFNSNFISLAKFDDKSNKYEFFNSETGQS
BC5098(199-310)                ILGG--TLWHGTKVLDEAGNDVTQFNSNFISLAKFDDKSNKYEFFNSETGQS
RBTH_06214(199-310)            ILGG--TLWHGTKILDEAGNDVTQFNSNFISLAKFDDKSNKYEFFNSETGQS
                                 .     * ** : *   *::*  *  ::   *:*     ****: .** :
consensus/80%                  ILut..T.WpGT+VhDcstNDlTp.NtNhIuLAKaDtpos+YEFFshpTGpS

Secondary structure                  EEE         EEEEE       EEEE        EEE
EF0374(62-172)               RGDEGTFFVTD---DGEKRILISDTQN-YQAVVDLTEVTKDKFTYKRM
EF0375(58-168)               RGDEGTFFVTG---DGTKRILISRTQN-YQAVVDLTEVSKDKFTYKRL
EF0376(59-172)               TGDEGTFFMTAGITDVSRLVIISETKN-YQGVYPLRTLYQDTFTYRQM
BA5326(58-168)               RGDKGTFFITN---DGKKRILISESMK-YQAVVDMTKLNKNVFTYKRM
BCZK4809(58-168)             RGDKGTFFITN---DGKKRILISESMK-YQAVVDMTKLNKNVFTYKRM
BT9727_4791(58-168)          RGDKGTFFITN---DGKKRILISESMK-YQAVVDMTKLNKNIFTYKRM
BC5098(58-168)               RGDKGTFFVTN---DGKKRILISESMK-YQAVIDMTKLNKNVFTYKRM
RBTH_06214(58-168)           RGDKGTFFVTN---DGKKRILISESMK-YQAVVDMTKLNKNVFTYKRM
BA3695(247-357)              RNDSGTFFITN---DGKKRVLISETQN-YQAVVELTQLDKEKFTYKRM
Bant_01004347(247-357)       RNDSGTFFITN---DGKKRVLISETQN-YQAVVELTQLDKEKFTYKRM
BT9727_3386(247-357)         RNDSGTFFITN---DGKKRVLISETQN-YQAVVELTQLDKEKFTYKRM
BCZK3337(229-339)            RNDSGTFFITN---DGKKRVLISETQN-YQAVVELTQLDKEKFTYKRM
BCE_G9241_3590(229-339)      RNDSGTFFITN---DGKKRVLISETQN-YQAVVELTQLDKEKFTYKRM
EF0376(223-336)              RGDYGYFKVGN---QNKFRAHVSIGTNRYGAVLELTELNDNRFTYTRM
EF0375(199-310)              RGDFGYFQVVD---NNKIRAHVSIGTNRYGAALELTELNNDRFTYTRM
EF0374(203-314)              RGDFGYFQVID---NNKIRAHVSIGDNKYGAALELTELNDKRFTYTRM
BA3695(388-499)              RGDYGFFDVVH---DNKIRAHVSLGNNKYGAVLELTELNKEKFTYTRM
BT9727_3386(388-499)         RGDYGFFDVVH---DNKIRAHVSLGNNKYGAVLELTELNKEKFTYTRM
Bant_01004347(388-499)       RGDYGFFDVVH---DNKIRAHVSLGNNKYGAVLELTELNKEKFTYTRM
BCZK3337(370-481)            RGDYGFFDVVH---GNKIRAHVSLGNNKYGAVLELTELNKAKFTYTRM
BCE_G9241_3590(370-481)      RGDYGFFDVVH---GNKIRAHASLGNNKYGAVLELTELNKEKFTYTRI
BA5326(199-310)              RGDYGYFDVLH---ENKIRAHVSIGNNKYGAALELTELNKNKFTYKRT
BCZK4809(199-310)            RGDYGYFDVLH---ENKIRAHVSIGNNKYGAALELTELNKNKFTYKRT
BT9727_4791(199-310)         RGDYGYFDVLH---ENKIRAHVSIGNNKYGAALELTELNKNKFTYKRT
BC5098(199-310)              RGDYGYFDVVH---ENKIRAHVSIGNNKYGAALELTELNKNKFTYKRT
RBTH_06214(199-310)          RGDYGYFDVVH---ENKIRAHVSIGNNKYGAALELTELNKNKFTYKRT
                               .* * * :             *   : * ..  :  : .  *** :
consensus/80%                RGD.GhF.lsp...-sKhRhhlS.spN.YtAsl-LTpLsKppFTYpRh
```

```
Secondary structure                    EEEEEE
EF0374(62-172)                 GKDKDGKDVEVFVEHIP 111
EF0375(58-168)                 GKDKLGNDVEVYVEHIP 111
EF0376(59-172)                 GKDKNGNDIEVFVENKA 114
BA5326(58-168)                 GKDANGNDVEVFVEHVP 111
BCZK4809(58-168)               GKDANGNDVEVFVEHVP 111
BT9727_4791(58-168)            GKDANGNDVEVFVEHVP 111
BC5098(58-168)                 GKDANGKDVEVFVEHVP 111
RBTH_06214(58-168)             GKDANGKDVEVFVEHVP 111
BA3695(247-357)                GKDAKRNDVEVFVEHIP 111
Bant_01004347(247-357)         GKDAKRNDVEVFVEHIP 111
BT9727_3386(247-357)           GKDAKGNDVEVFVEHIP 111
BCZK3337(229-339)              GKDAKGNDVEVFVEHVP 111
BCE_G9241_3590(229-339)        GKDVKGNDVEVFVEHIP 111
EF0376(223-336)                GKDNEGNDIQVYVEHEP 114
EF0375(199-310)                GKDNAGNDIQVFVEHEP 112
EF0374(203-314)                GKDNNGKEIKVFVEHEP 112
BA3695(388-499)                GKDANGKDIKIFVEHEP 112
BT9727_3386(388-499)           GKDANGKDIKIFVEHEP 112
Bant_01004347(388-499)         GKDANGKDIKIFVEHEP 112
BCZK3337(370-481)              GKDANGKDIKIFVEHEP 112
BCE_G9241_3590(370-481)        GKDANGKDIKIFVEHEP 112
BA5326(199-310)                GKDQAGNDITIFVEHEP 112
BCZK4809(199-310)              GKDQAGNDITIFVEHEP 112
BT9727_4791(199-310)           GKDQAGNDITIFVEHEP 112
BC5098(199-310)                GKDQAGKDITIFVEHEP 112
RBTH_06214(199-310)            GKDQAGKDITIFVEHEP 112
                               ***   ::: ::**: .
consensus/80%                  GKDttGpDlplFVEH.P
```

BA3695 is homologous to proteins GBAA3695 from *B. anthracis* str. "Ames Ancestor" and BAS342 from *B. anthracis* str. *Sterne*. BA5326 is homologous to proteins GBAA5326 from *B. anthracis* str. "Ames Ancestor," BAS4948 from *B. anthracis* str. Sterne and Bant_01000199 from *B. anthracis* str. A2012.

Figure 4.1d: Multiple sequence alignment of 109 amino acid residue IMxxH domain.

```
Secondary structure                    HHHHHHHHHHHHHHHHHH          HHHHHHHHHHH
BCE_G9241_1042_1(21-129)          ERSLNEIRFWSRIMKEHSLFLRLGFRCEDTQLIEEANQFYRLF
BCZK0933_1(21-129)                ERSLNEIRFWSRIMKEHSLFLRLGFRCEDTQLIEEANQFYRLF
BT9727_0941_1(21-129)             ERSLNEIRFWSRIMKEHSLFLRLGFRCEDTQLIEEANQFYRLF
BA1021_1(4-112)                   ERSLNEIRFWSRIMKEHSLFLRLGFRCEDTQLIEEANQFYRLF
BAS0955_1(21-129)                 ERSLNEIRFWSRIMKEHSLFLRLGFRCEDTQLIEEANQFYRLF
RBTH_03050_1(21-129)              ERSLNEIRFWSRIMKEHSFFLRLGFRCEDTQLIEEANQFYRLF
BC1029_1(21-129)                  ERSLNEIRFWSRIMKEHSFFLRLGFRCEDTQLIEEANQFYRLF
BcerKBAB4DRAFT_3543_1(21-129)     ERSLNEIRFWSRIMKEHSFFLRLGFRCEDTQLIEEANQFYRLF
Bcer98DRAFT_1038_1(42-147)        EKSLTENRFWLRIMKEHALFLGEGFNRKDTNLIQQVDQFFHLF
CTC02189(189-294)                 RYAYEQETFWNRIMAEHAKFIRGLLDPTEDALIDTANNFGKEF
CbeiDRAFT_3331(190-295)           REAYEQEAFWNRIMAEHSKFIRGLLDPTEDELINTANNFGHQF
ClosDRAFT_1658(189-294)           KEIYEQELFWNRIMAEHSKFIRGLLDPTEDELIHIANDFAKEF
CtheDRAFT_1311(189-294)           KEAYELQFFWNRQMAEHAKFIRGLLDPTENDLINQANDFGNEF
CdifQ_02001573(138-241)           KNAKEIELFWDHIMMEHALFMRGLLDPSEGELINTSNDFAIKF
CD1511(189-291)                   KNAKEIELFWDHIMMEHALFMRGLLDPSEGELINTSNDFAIKF
CPE0158_2(188-291)                VNISKTEAFWNEIMMEHSLFIRGLLDPSEYELINTAHEFAFEF
CPF_0149(188-291)                 VNISKTEAFWNEIMMEHSLFIRGLLDPSEYELINTAHEFAFEF
CphyDRAFT_3436(189-292)           EDLKDDELFWNQIMMEHALFIRGLLDPTENDLIMQADDFASVY
DhafDRAFT_0725_2(197-302)         CHMVEMQMFWDHIMKEHAEVISHLLDPKEKAMITRADHFAQAY
BCZK0933_2(149-260)               DAIIKENVFFLRIMADHAKFIGHLLDPSERKLVDTARNFSNDF
BT9727_0941_2(149-260)            DAIIKENVFFLRIMADHAKFIGHLLDPSERKLVDTARNFSNDF
BA1021_2(132-243)                 DAIIKENVFFLRIMADHAKFIGHLLDPSERKLVDTARNFSNDF
BAS0955_2(149-260)                DAIIKENVFFLRIMADHAKFIGHLLDPSERKLVDTARNFSNDF
BCE_G9241_1042_2(149-260)         DAIIKENVFFLRIMADHAKFIGHLLDPSERKLVDTARNFSNDF
BC1029_2(149-260)                 DAIIKENVFFLRIMADHAKFIGHLLDPSERKLVDTARNFSNDF
RBTH_03050_2(149-260)             DAIIKENVFFLRIMADHAKFIGHLLDPSERKLVDTARNFSNDF
BcerKBAB4DRAFT_3543_2(149-260)    DAIIKENVFFLRIMADHAKFIGHLLDPSERKLVDTARNFSNDF
BcerKBAB4DRAFT_0307(35-147)       DAIISENVFWLRIMMEHSRFIGSLLDQSERNLVHTALKFGDDF
Bcer98DRAFT_1038_2(167-279)       DAIISENVFWLRIMMEHSRFIASLLDQSERNLVHTALKFGDDF
CAC3450_1(190-295)                QGIIRQEIFWNDIMEDHAEFIRGYLDPSQTSLFNTANNFVRRF
CPE0158_1(9-119)                  TSSLELHLFFMRVMKEHAIFLEAGLGPKNSKLAKELDKCKGNL
DhafDRAFT_0725_1(12-122)          RESLELHLFWARIIKEHLIFLESGFMCKDADWMQEADALKCSF
CAC3450_2(9-121)                  RLSLELNLFFLRIVKEHNVIAGASLPPKYAPTLMEILAVNKKL
AmetDRAFT_1908_1(11-115)          NVALFEHQFWLQVLGDHARFILNALSPEEREEIQRAQYFIHIF
AmetDRAFT_1908_2(133-245)         TQPIHYHMVWLLDAAGHSAGIMGDLDMVEKELIRKSGKFTQRF
                                            .:      *       :
consensus/80%                     ct.hp..hFa.+IMt-HuhFlthhhcsp-ppLlppAppF.p.F
```

Contd…..

```
Secondary structure                HHHHHHHH                    HHHHHHH  HHHHHHHHH
BCE_G9241_1042_1(21-129)           EHIEQIAHSYTNETDPEQ-----IKRFNAEVQQAATNIWGFKRKILG
BCZK0933_1(21-129)                 EHIEQIAHSYTNETDPEQ-----IKRFNAEVQQAATNIWGFKRKILG
BT9727_0941_1(21-129)              EHIEQIAHSYTNETDPEQ-----IKRFNAEVQQAATNIWGFKRKILG
BA1021_1(4-112)                    EHIEQIAHSYTNETDPEQ-----IKRFNAEVQQAATNIWGFKRKILG
BAS0955_1(21-129)                  EHIEQIAHSYTNETDPEQ-----IKRFNAEVQQAATNIWGFKRKILG
RBTH_03050_1(21-129)               EHIEQIAHSYTNETDPEQ-----IKRFNAEVQQAATNIWGFKRKILG
BC1029_1(21-129)                   EHIEQIAHSYTNETDPEQ-----IKRFNAEVQQAATNIWGFKRKILG
BcerKBAB4DRAFT_3543_1(21-129)      EHIEQIAYSYTNETDPGQ-----IKRFNSEVQQAATNIWGFKRKILG
Bcer98DRAFT_1038_1(42-147)         DRHLQKAFSIP--QTVQA-----VRQLNEESIQLVYAFRNYKRNLLI
CTC02189(189-294)                  DELTR---EAKRAMYKTM----PISKVTNRSLRATRRIRNFKKQGTE
CbeiDRAFT_3331(190-295)            DILTR---EARAAMNKSI----PISKVTDESLEATKSIRNFKAQGTQ
ClosDRAFT_1658(189-294)            DALTA---AVEEAIEKCL----PIDKITDKSLEATKEVRNFNTQGTE
CtheDRAFT_1311(189-294)            DQLTA---EAKAAMDATS----PMAKVTDESLKATEDFRNFKAQGTQ
CdifQ_02001573(138-241)            NELIE---KTN--EMTDS----NIKNITEETLNETVEFKDFKEAGAS
CD1511(189-291)                    NELIE---KTN--EMTDS----NIKNITEETLNETVEFKDFKEAGAS
CPE0158_2(188-291)                 NELIQ---QLN--NVTNV----TIDNVTHEILKETTRLRDFKEEGTK
CPF_0149(188-291)                  NELIQ---QLN--NVTNV----TIDNVTHETLKETTRLRDFKEEGTK
CphyDRAFT_3436(189-292)            ADLLD---EAS--TMTER----TMGDLTCRTLEETIKYRDFKLAGTK
DhafDRAFT_0725_2(197-302)          EQLLN---QLGNGTVPDQ----SFRRITSETIRVTGEFKDFKAAGTD
BCZK0933_2(149-260)                DALMYQAIDLESMKPQSQ-TVPLLDQFLDQNRVSVASLRDFKKTARD
BT9727_0941_2(149-260)             DALMYQAIDLESMKPQSQ-TVPLLDQFLDQNRVSVASLRDFKKTARD
BA1021_2(132-243)                  DALMYQAIDLESMKPQSQ-TVPLLDQFLDQNRVSVASLRDFKKTARD
BAS0955_2(149-260)                 DALMYQAIDLESMKPQSQ-TVPLLDQFLDQNRVSVASLRDFKKTARD
BCE_G9241_1042_2(149-260)          DALMYQAIDLESMKPQSQ-TVPLLDQFLDQNRVSVASLRDFKKTARD
BC1029_2(149-260)                  DELMYQAIDLESMKPQSQ-TAPLLDQFLDQNRVSVASLRDFKKTARD
RBTH_03050_2(149-260)              DELMYQAIDLESMKPQSQ-TVPLLDQFLDQNRVSVTSLRDFKKTARD
BcerKBAB4DRAFT_3543_2(149-260)     DELMYQAIDLESMKPQSQ-TVPLLDQFLDQNRVSVTSLRDFKKTARD
BcerKBAB4DRAFT_0307(35-147)        EILLNQARDVESMLYQKEPTYPIIGKMNKDSENATVELRNFKKAGLE
Bcer98DRAFT_1038_2(167-279)        EVLLSQARDVESMLYQKQPTYPIIGKMNKDSENATVELRNFKKAGLE
CAC3450_1(190-295)                 DDIEN---ATESLTNNPS----NLNNITRNIYSLVTEFRNFKSTATK
CPE0158_1(9-119)                   EKLLFDVVKLSKGRVRQSIVD-SGEVFTDYTLETEKKTEHYTGININ
DhafDRAFT_0725_1(12-122)           EEILHEANCLADGKVGIEVMK-SGELFTNKTLKAEQKTQELTCIPIN
CAC3450_2(9-121)                   DMLLSKTVALSKGNISREAMN-SSTLITPLTLPSEKVTSALTGVPIN
AmetDRAFT_1908_1(11-115)           DQLLE----ESRKSPRGS----ALSKLTDQAYGCAQEIRTFKLHLIK
AmetDRAFT_1908_2(133-245)          EEFYIKAVEIAGYTRTTLDQFPAFTRFNYQVEGELLLFKKFLRELEA

consensus/80%                      -l.....phpt.p..pp.....lpph.tps..tstphhsFKpthhth
```

BAS0955 is homologous to proteins BT9727_0941 from *B. thuringiensis* serovar konkukian str. 97-27, BCZK0933 from *B. cereus* E33L, and BCE_G9241_1042 from *B. cereus* G9241. BA1021 is homologous to protein GBAA1021 from *B. anthracis* str. “Ames Ancestor.” BA0807 is homologous to proteins GBAA0807 from *B. anthracis* str. “Ames Ancestor” and BAS0770 from *B. anthracis* str. Sterne.

Contd…..

```
Secondary structure
BCE_G9241_1042_1(21-129)          LILTCKLPGQNNFPLLVDHTSREA 109
BCZK0933_1(21-129)                LILTCKLPGQNNFPLLVDHTSREA 109
BT9727_0941_1(21-129)             LILTCKLPGQNNFPLLVDHTSREA 109
BA1021_1(4-112)                   LILTCKLPGQNNFPLLVDHTSREA 109
BAS0955_1(21-129)                 LILTCKLPGQNNFPLLVDHTSREA 109
RBTH_03050_1(21-129)              LILTCKLPGQNNFPLLVDHTSREA 109
BC1029_1(21-129)                  LILTCKLPGQNNFPLLVDHTSREA 109
BcerKBAB4DRAFT_3543_1(21-129)     LILTCKLPGQNNFPLLVDHTSREA 109
Bcer98DRAFT_1038_1(42-147)        LIINCKVSGFN-FPLLVDHIAREA 106
CTC02189(189-294)                 GILDCKIRSII-IPLLADHTLREA 106
CbeiDRAFT_3331(190-295)           GLVECKIKSII-IPLLGDHTLREA 106
ClosDRAFT_1658(189-294)           GLLDCKIRSII-IPLLGDHVLRES 106
CtheDRAFT_1311(189-294)           AILECKVKSII-IPLLGDHVLREA 106
CdifQ_02001573(138-241)           GIEQCKIKSII-LPLLADHVLREA 104
CD1511(189-291)                   GIEQCKIKSII-LPLLADHVLREA 104
CPE0158_2(188-291)                GIMNCNIKSLI-LPLLSDHVLREA 104
CPF_0149(188-291)                 GIMNCNIKSLI-LPLLSDHVLREA 104
CphyDRAFT_3436(189-292)           GINDCEIRSII-LPLLADHVLREA 104
DhafDRAFT_0725_2(197-302)         AILCCQLRSLI-LPLLADHVLREA 106
BCZK0933_2(149-260)               LIEQCKIKSII-HPLLADHVFREA 112
BT9727_0941_2(149-260)            LIEQCKIKSII-HPLLADHVFREA 112
BA1021_2(132-243)                 LIEQCKIKSII-HPLLADHVFREA 112
BAS0955_2(149-260)                LIEQCKIKSII-HPLLADHVFREA 112
BCE_G9241_1042_2(149-260)         LIEQCKIKSII-HPLLADHVFREA 112
BC1029_2(149-260)                 LIEQCKIKSII-HPLLADHVFREA 112
RBTH_03050_2(149-260)             LIEQCKIKSII-HPLLADHVFREA 112
BcerKBAB4DRAFT_3543_2(149-260)    LIEQCKIKSII-HPLLADHVFREA 112
BcerKBAB4DRAFT_0307(35-147)       LIQTCQIRSVI-NPLLADHVTREA 113
Bcer98DRAFT_1038_2(167-279)       LIQTCQIRNVI-NPLLADHVVREA 113
CAC3450_1(190-295)                GLLACKIKAIM-APLLADHVTREA 106
CPE0158_1(9-119)                  SKITTMEKDLMC--APKKGIDSKV 111
DhafDRAFT_0725_1(12-122)          SQLTVETMSLHP--YMGVGMGMVP 111
CAC3450_2(9-121)                  TAITSKEISLGYRDYYRTGINMVT 113
AmetDRAFT_1908_1(11-115)          RHLVGKIEIGL-PPTFLNHMVNEV 105
AmetDRAFT_1908_2(133-245)         LELNQKVLGTL-SALMLDHMAREE 113

consensus/80%                     l.pCcl.u....hPLLsDHs.REA
```

Figure 4.1e: Multiple sequence alignment of 103 amino acid residue VxxT domain.

```
Secondary structure                    EE                  HHHHHHHHHH
BT9727_4219_1(67-169)          VDKNGYVVPQTLAIPTPKANE----VIQQTLEYLVKDGPVTNLLPN-GF
BCZK4235_1(67-169)             VDKNGYVVPQTLAIPTPKANE----VIQQTLEYLVKDGPVTNLLPN-GF
BA4716_1(67-169)               VDKNGYVVPQTLAIPTPKANE----VIQQTLEYLVKDGPVTNLLPN-GF
BCE4587_1(67-169)              VDKNGYVVPQTLAIPTPKANE----VIQQTLEYLVKDGPVTNLLPN-GF
RBTH_05210_1(67-169)           VDKNGYVVPQTLAIPTPKANE----TVKQTLEYLVKDGPVTNLLPN-GF
ABF83609_1(67-169)             VDKNGYVVPQTLAIPTPKANE----TVKQTLEYLVKDGPVTNLLPN-GF
BC4495_1(67-169)               VDKNGYVVPQTLAIPTPKANE----TVKQTLEYLVKDGPVTNLLPN-GF
BcerKBAB4DRAFT_4089_1(67-169)  VDKNGYVVPQTIAMPTPKANE----VVQQTLEYLVKDGPVTNLLPN-GF
Bcer98DRAFT_3179_1(67-169)     VDKNGYVVPQTLALPIPKQSE----VVKQTLEYLVKDGPVENILPN-GF
gerM(84-184)                   IDKNGYVVAQTLPLPKSES------TAKQALEYLVQGGPVSEILPN-GF
BSU28380_1(84-184)             IDKNGYVVAQTLPLPKSES------TAKQALEYLVQGGPVSEILPN-GF
BL00314_1(87-187)              IDKNGYVTAQTLPLPKQEG------TAKQALEYLVEGGPVSNILPN-GF
GK2667_1(76-177)               IDKNGFVVPQTVELPKTQA------VAKQVLEYLVEDGPVSEMLPN-GF
BH3070_1(87-186)               LDENGMVVPQTLPLPKSDG------VLKQSLEYLVEGGPVTNLLPN-GF
OB2107_1(69-172)               LDANGMVASQTLELPVPDTNE----VAAQVLEHLVKGGPVTPLLPN-GF
B14911_06091_1(82-181)         VDKNGYVVPQTLTLPKTES------VATQALEYLMQNGPVTDMLPN-DF
ABC2653_1(99-200)              IDSNGLVVPQTLTLPKTDS------VMKQALEYLVEGGPINDILPN-GF
SwolDRAFT_2302(77-173)         ADKEELVMERR-EITRTEG------IARSTLQELLK-GPDN---P--AY
Moth_0516(72-172)              DSSGNYLVAEKRSIPAVEG------IARATIEELIKGPAPDSK-----L
MothDRAFT_0979(72-172)         DSSGNYLVAEKRSIPAVEG------IARATIEELIKGPAPDSK-----L
CtheDRAFT_0840(63-168)         NEDNSKLKLEIRYIPVSETTKSVNHLAEIIVNELIKGPKVAG------L
AmetDRAFT_1640_1(62-164)       RDDKGLLIPVMRRIPWQEG------IAKAALEQLVDQPVLRDDLATIGL
GAA01614(67-167)               TGSDAYLVREVHQVPFTRE------VAKAALEELINTAPSTPG-----A
BCE4587_2(220-319)             NNKQQYYVPVTRRVVEGKE-----NDYAAIVDELVKGPIHQS------L
BA4716_2(220-319)              NNKQQYYVPVTRRVVEGKE-----NDYAAIVDELVKGPIHQS------L
BT9727_4219_2(220-319)         NNKQQYYVPVTRRVVEGKE-----NDYAAIVDELVKGPIHQS------L
BCZK4235_2(220-319)            NNKQQYYVPVTRRVVEGKE-----NDYAAIVDELVKGPIHQS------L
ABF83609_2(20-319)             NNKQQYYVPVTRRVAEGKE-----NDYAAIIDELVKGPIHQS------L
BC4495_2(220-319)              NNKQQYYVPVTRRVAEGKE-----NDYATIIDELVKGPIHQS------L
RBTH_05210_2(220-310)          NNKQQYYVPVTRRVAEGKE-----NDYAAIIDELVKGPIHQS------L
BcerKBAB4DRAFT_4089_2(220-319) NNKQQYYVPVTRRVAEGKE-----NDYSAIVDELVKGPIQGS------L
Bcer98DRAFT_3179_2(219-318)    NNKRQYYVPVTRRVAEEKE-----NEVETIINELVKGPSHSS------L
BSU28380_2(234-336)            NEDSEYYVPVTKRIDNSEK-----DDITAAINELAKGPSKVSG-----L
BL00314_2(237-339)             SDKGTYYVPVTKRTSAKEK-----DQVTAAIKELTEGPDNKSG-----L
GK2667_2(227-327)              QGNSTYYVPVTRRVSNKEK-----DDIAAAVNELIQGPEQGSG-----L
B14911_06091_2(231-331)        EEGAYYYVPVTKRISAQED-----NQVEAVVKELVKGPSFTSN-----L
BH3070_2(236-335)              SGDQTYYVPVTRRVNVKD------NSFATAVEELLNGPMVTSP-----L
ABC2653_2(250-349)             NDEDTYYVPVTKRVENVD------NELEAAINELIDGPSLMTN-----L
OB2107_2(222-322)              QENNRYYVPVTQYIETNED-----EAIANIIKELIDGPGHQSK-----V
AmetDRAFT_1640_2(210-309)      NGEDDFFIPITRGLNVLKA------DTKSVLTALVEGAPVGSG-----L
                                                          :  * .
consensus/80%                  .scptYhVs.Thtlstsct.......htthlc.Llcss.hps......h
```

Contd…..

```
Secondary structure                          EEEEEE                HHH
BT9727_4219_1(67-169)          RAVIPANTSMT--LDLKKDGTAVIDFSKEMKNYA----KEEERQIV
BCZK4235_1(67-169)             RAVIPANTSMT--LDLKKDGTAVIDFSKEMKNYA----KEEERQIV
BA4716_1(67-169)               RAVIPANTSMT--LDLKKDGTAVIDFSKEMKNYA----KEEERQIV
BCE4587_1(67-169)              RAVIPANTSMT--LNLKKDGTAVIDFSKEMKNYA----KEEERQIV
RBTH_05210_1(67-169)           RAVIPANTTMT--LDLKKDGTAVIDFSKEMKNYA----KEEERQIV
ABF83609_1(67-169)             RAVIPANTTMT--LDLKKDGTAVIDFSKEMKNYA----KEEERQIV
BC4495_1(67-169)               RAVIPANTTMT--LDLKKDGTAVIDFSKEMKNYA----KEEERQIV
BcerKBAB4DRAFT_4089_1(67-169)  RAVLPANTTMT--LNLKKGGTAVIDFSKEMKNYS----KEEERQIV
Bcer98DRAFT_3179_1(67-169)     RAVLPADTTMT--VDLKKDGTAVIDFSKEMQNYK----KEEERQIV
gerM(84-184)                   RAVLPADTTVN--VDIKKDGTAIADFSNEFKNYK----KEDEQKIV
BSU28380_1(84-184)             RAVLPADTTVN--VDIKKDGTAIADFSNEFKNYK----KEDEQKIV
BL00314_1(87-187)              RAVLPADTTVN--VDIKEDGTAIADFSNEFKNYK----AEDEQKIV
GK2667_1(76-177)               RAVIPAGTTVL-GTKLEKDGTLIADFSPEFKNYK----PEDEKRIL
BH3070_1(87-186)               QAVLPPDTEMS--VNL-EDGVAVVDFSKEFTEYD----GEKEQQIL
OB2107_1(69-172)               QAVLPEGTEVL-GVNLQEDGTIIVDLSEEFTQYE----ENQEVQIL
B14911_06091_1(82-181)         RAVLPADTKIS-VN--VKDKVATVDFSKEFGDYQ----AEDEEKIL
ABC2653_1(99-200)              RAVLPAGTEVD-IDHLKEEKLAIVNFSSEFNDYN----LADEKQIF
SwolDRAFT_2302(77-173)         RNVFPEGTRLL-DINLKPDGTCILDFSSELRRLEN---EVEEKQML
Moth_0516(72-172)              LPTIPKGTVLK-DINIRPDGLARVDFSKELVANHS-GGSLGESLTV
MothDRAFT_0979(72-172)         LPTIPKGTVLK-DINIRPDGLARVDFSKELVANHS-GGSLGESLTV
CtheDRAFT_0840(63-168)         KPTIPEGTKLRSAIKIEGD-VAIVDFTKEFRDNHP-GGKAEERMTI
AmetDRAFT_1640_1(62-164)       LPVLPPGTEVI-GISINEG-LSKVDFNEQLLAYQS---EIDENAIV
GAA01614(67-167)               VRVLPPATKIR-GISIKDG-LATVDFSRDVLRANT-G-ASGEALGI
BCE4587_2(220-319)             LNDFNPGVKLI-TNPKLQDGNLTLNFNENIFINP--DKNMISNYVL
BA4716_2(220-319)              LNDFNPGVKLI-TNPKLQDGNLTLNFNENIFINP--DKNMISNYVL
BT9727_4219_2(220-319)         LNDFNPGVKLI-TNPKLQDGNLTLNFNENIFINP--DKNMISNYVL
BCZK4235_2(220-319)            LNDFNPGVKLI-TNPKLQDGNLTLNFNENIFINP--DKNMISNYVL
ABF83609_2(20-319)             LNDFNPGVKLI-TNPKLQDGNLTLNFNENIFVNP--DKNMISNYVL
BC4495_2(220-319)              LNDFNPGVKLI-TNPKLQDGNLTLNFNENIFVNP--DKNMISNYVL
RBTH_05210_2(220-310)          LNDFNPGVKLI-TNPKLQDGNLTLNFNENIFVNP--DKNMISNYVL
BcerKBAB4DRAFT_4089_2(220-319) LNDFNPGAKLI-TNPKVENGNITLNFNENIFVNP--DKNMISNYVL
Bcer98DRAFT_3179_2(219-318)    LNDFNPGVKLV-SEPKIQDGKVTLNFNENIYANK--DKNMISNYVL
BSU28380_2(234-336)            LTDFSEDVKLV-SKPKIKDGRVTLDFNQSIFGSADEKTKMISSEVL
BL00314_2(237-339)             LSDFQGDVKLE-NKPKIEDGHVTLDFNEAIYGSADGQKKVISDEVL
GK2667_2(227-327)              VGVFQPDAKLV-DAPKYEDGKVTLNFNEGIYGSN--KKNVISDVVL
B14911_06091_2(231-331)        FTDFMPEVELL-GDPKIENGLATLDFNESVYGSF--EEKIISQHLL
BH3070_2(236-335)              VTDFRNGVELL-DEPKYENGVVTLNFNEALLSQM--QATAVSDEII
ABC2653_2(250-349)             LTEMSGDVELL-NEPKLQNGEVVLDFNEAIQSAN--EGSAIPTSVL
OB2107_2(222-322)              VNVFNPEAGLA-SEPTLNNGILEVVFNKEILADS--EQGIIADEVM
AmetDRAFT_1640_2(210-309)      HSEIPYGASIN--DVYVRDGIAYIDFTEEIRNVP--VNEKHQQSLV
                                :   . :              :.  .            .
consensus/80%                  hsshssssphh...shhp-G.hhlsFscphhs......p..pp.ll
```

Contd…..

```
Secondary structure                HHHHHHH         EEEE
BT9727_4219_1(67-169)              ESIAWTLTQFK-EVKQVQFQ 103
BCZK4235_1(67-169)                 ESIAWTLTQFK-EVKQVQFQ 103
BA4716_1(67-169)                   ESIAWTLTQFK-EVKQVQFQ 103
BCE4587_1(67-169)                  ESIAWTLTQFK-EIKQVQFQ 103
RBTH_05210_1(67-169)               ESIAWTLTQFT-EIKQVQFQ 103
ABF83609_1(67-169)                 ESIAWTLTQFT-EIKQVQFQ 103
BC4495_1(67-169)                   ESIAWTLTQFT-EIKQVQFQ 103
BcerKBAB4DRAFT_4089_1(67-169)      ESVAWTLTQFT-EIKQVQFQ 103
Bcer98DRAFT_3179_1(67-169)         ESVAWTLTQFK-DIKQVKFQ 103
gerM(84-184)                       QSVTWTLTQFS-SIDKVKLR 101
BSU28380_1(84-184)                 QSVTWTLTQFS-SIDKVKLR 101
BL00314_1(87-187)                  QAITWTLTQFN-SIDKVKLR 101
GK2667_1(76-177)                   QSITWTLTQFD-NIKRVKIR 102
BH3070_1(87-186)                   QSITWTLTQFE-NVEKVKLQ 100
OB2107_1(69-172)                   ESVTHTLTQFE-SVHKVKLR 104
B14911_06091_1(82-181)             ESITWTLTQFD-SIEKVKLQ 100
ABC2653_1(99-200)                  EAVTWTLTQFP-DVEEVKVE 102
SwolDRAFT_2302(77-173)             DAVCQTLAQFP-AVKQLVFM 97
Moth_0516(72-172)                  YSIVNTLTQFP-TIKQVQFL 101
MothDRAFT_0979(72-172)             YSIVNTLTQFP-TIKQVQFL 101
CtheDRAFT_0840(63-168)             YSVVNSLTELK-EINKVKFL 106
AmetDRAFT_1640_1(62-164)           KSIVYTLTEFD-SIDQVQIM 103
GAA01614(67-167)                   QSIVNTLTEFP-EVQKVSFL 99
BCE4587_2(220-319)                 KSLVLSLTEKK-GVKSVSIE 100
BA4716_2(220-319)                  KSLVLSLTEKK-GVKSVSIE 100
BT9727_4219_2(220-319)             KSLVLSLTEKK-GVKSVSIE 100
BCZK4235_2(220-319)                KSLVLSLTEKK-GVKSVSIE 100
ABF83609_2(20-319)                 KSLVLSLTEKK-GVKNISIE 100
BC4495_2(220-319)                  KSLVLSLTEKK-GVKNISIE 100
RBTH_05210_2(220-310)              KSLVLSLTEKK-GVKNVSIE 100
BcerKBAB4DRAFT_4089_2(220-319)     KSLVLSLTEKQ-GVKNVSIE 100
Bcer98DRAFT_3179_2(219-318)        QSLVLSLTEKQ-GVKNVSVE 100
BSU28380_2(234-336)                NSIVLTLTEQP-DVKSVSVK 103
BL00314_2(237-339)                 NSIVLTLTELP-DVKSVSVT 103
GK2667_2(227-327)                  NSLVLSLTEQK-GVESVAIT 101
B14911_06091_2(231-331)            NSLVLSLTEQK-GIESVAVT 101
BH3070_2(236-335)                  NMLALTLTEQD-GVEKVAIQ 100
ABC2653_2(250-349)                 ESLALTLTEQG-GIEKVSIQ 100
OB2107_2(222-322)                  ETMVRTLTEQP-NIDAVDVK 101
AmetDRAFT_1640_2(210-309)          YELGLTLREVEPSIHQVRIL 100
                                     :  :* :    :. : .
consensus/80%                      pSlshoLTpht.tlcpVphp
```

BA4716 is homologous to proteins GBAA4716 from *B. anthracis* str. "Ames Ancestor," BAS4378 from *B. anthracis* str. Sterne, and Bant_ 01005366 from *B. anthracis* str. A2012. BT9727_4219 is homologous to protein BCZK4235 from *B. cereus* E33L. BA4716 is homologous to protein BL02986 from *B. licheniformis* ATCC 14580.

Figure 4.1f: Multiple sequence alignment of 84 amino acid residue ExW domain.

```
Secondary structure                        EEEEEE              EEEEE
BC4088_1(47-130)                  IKPGEKTEVQALVTQGKEKVTDADDVKFEIWKDGD--EKHEMLDGKHK
RBTH_02670_1(47-130)              IKPGEKTEVQALVTQGKEKXTDADDVKFEIWKDGD--EKHEMLDGKHK
BCE_G9241_4093_1(45-128)          IKPGEKTEVQALVTQGKERVTDADDVKFEIWKDGD--EKHEMLDGKHK
BA4310_1(45-128)                  IKPGEKTEVQALVTQGKEKVTDADDVKFEVWKAGD--EKHEMLEGKHK
BT9727_3829_1(45-128)             IKPGEKTEVQALVTQGKEKVTDADDVKFEVWKAGD--EKHEMLEGKHK
Bant_01004966_1(51-134)           IKPGEKTEVQALVTQGKEKVTDADDVKFEVWKAGD--EKHEMLEGKHK
BCE4157_1(45-128)                 IKPGEKTEVQALVTQGKEKVTDADDVKFEIWKAGD--EKHEMLEGKHK
BCZK3845_1(45-128)                IKPGEKTEVQALVTQGKEKVTDADDVKFEIWKAGD--EKHEMLEGKHK
BcerKBAB4DRAFT_2040_1(46-128)     IKPGEKTEVQALVTQGKEKVTDADDVKFEIWKAGD--EKHEMLNAKHK
GK0969(45-128)                    IDLNKPTKLACVVTYGGEKVDDANEVKFEVWKHGS--DEREMLEAKHD
BL05305(45-129)                   AAKNEKAVIKATVLYGEEPVADADEVEFECWKAGSK-EDSELIKAKNE
BSU30660(44-127)                  VNPGESAAYEAAVSYGDEAVTDADEVEFEVWKEGEK-DASQMFKVKQE
OB2488(50-134)                    VETGETIDLTAHVTYGDAPVEDADEVIFEVWTQGNS-DQSVELEGKHQ
B14911_05359_1(53-137)            VELNEEITLSVEVVQGEEAVEDADEVKFEIWQEGNQ-EESEMLPAEHT
BH0678_1(45-129)                  LASGENMTFDVLVTQNEAPVEDAREVIVEFWQEGAK-EESDMIESTNE
ABC0230(45-129)                   IEIGEEILLSVQLAQGEVQVEDADEVVFEVWKDQER-DNGTLQEATHQ
ABC4088(44-127)                   LEL-ENIVLEAKVMQGDEPVDDAEEVVFEVWPYDDR-EESEFHEASYA
BH0983(47-131)                    LIPNTPHELAIHVTQGDENVTDATDIQFEIWQGHDR-EQGELIEASHV
B14911_09907(34-118)              FAAGEDVPIRAVLTQNGEKVAGADYVHFEIWKRDGS-VHYPMEEAADE
ExigDRAFT_1796(51-135)            ADQEKQYRFGATLWQDQKAVKEAEYVHFEIWKADGT-LRYSMEPADET
BAA83944_1(46-130)                LVTDQEESLTVSLSHNGEILSKVDSLHVHIWKHDHT-VAYHFEQLETD
BH1853(46-130)                    LVTDQEESLTVSLSHNGEILSKVDSLHVHIWKHDHT-VAYHFEQLETD
OB3282(48-131)                    IEAKENTEVTFELSQNGESVSTLDDLSVTTWMVDSE-TTKQLVAENVG
BCE_G9241_4093_2(163-245)         IKANAESTMKVHLKQKE-EALTGAEVQLEIWKDGV--EKHEFIPAKEG
BC4088_2(165-247)                 IKANAESTMKVHLKQKE-EALTGAEVQLEIWKDGV--EKHEFIPAKEG
RBTH_02670_2(165-247)             IKANAESTMKVHLKQKE-EALAGAEVQLEIWKDGV--EKHEFIPAKEG
BcerKBAB4DRAFT_2040_2(158-240)    IKANAESTMKVHLKQKE-EALSGAEVQLEIWKDGV--EKHEFIPAKEG
BT9727_3829_2(163-245)            IKANAESTMKVHLKQKE-EALTGAEVQLEIWKDGV--EKHEFIPAKEG
Bant_01004966_2(169-251)          IKANAESTMKVHLKQKE-EALTGAEVQLEIWKDGV--EKHEFIPAKEG
BA4310_2(163-245)                 IKANAESTMKVHLKQKE-EALTGAEVQLEIWKDGV--EKHEFIPAKEG
BCE4157_2(163-245)                IKANAESTMKVHLKQKE-EALTGAEVQLEIWKDGV--EKHEFIPAKEG
BCZK3845_2(163-245)               IKANAESTMKVHLKQKE-EALTGAEVQLEIWKDGV--EKHEFIPAKEG
Bcer98DRAFT_3614(94-176)          VKANAESTLKAHVKQKE-EALTKAEVQFEIWKDGV--EKHTFITAKED
B14911_05359_2(187-271)           IHMKQAAGLDVQVDKKDGAPLEKALVKLEIMKEGK--DTPEWVNLKES
BH0678_2(159-242)                 IQAGEETTLLIVVEHKD-KPFTGGVLTLEVWQHED--EAHTWLDTEET
ExigDRAFT_0574(52-137)            KTMENQKVVFQATALENKKAVNLENVAFEVWKADEKEAVHQKFKAALK
                                                          : .
consensus/80%                     lp.stptphpshlpptc.ts.sus-VphElWKtss..-ppphh.ucpt
```

Contd…..

```
Secondary structure                    EEEEE       EEEEEEE             EE
BC4088_1(47-130)                     GKGVYAVEKTFETDGVYHIIAHTNARE-MHVMPEVKVAV 84
RBTH_02670_1(47-130)                 GKGVYAVEKTFETDGVYHIIAHTNARE-MHVMPEVKVAV 84
BCE_G9241_4093_1(45-128)             GKGVYAVEKTFETDGVYHIIPHTNARD-MHVMPEHKVAV 84
BA4310_1(45-128)                     GKGVYAVEKTFETDGVYHIIAHTNARE-MHVMPEVKVAV 84
BT9727_3829_1(45-128)                GKGVYAVEKTFETDGVYHIIAHTNARE-MHVMPEVKVAV 84
Bant_01004966_1(51-134)              GKGVYAVEKTFETDGVYHIIAHTNARE-MHVMPEVKVAV 84
BCE4157_1(45-128)                    GKGVYAVEKTFETDGVYHIIAHTNARE-MHVMPEVKVAV 84
BCZK3845_1(45-128)                   GKGVYAVEKTFETDGVYHIIAHTNARE-MHVMPEVKVAV 84
BcerKBAB4DRAFT_2040_1(46-128)        GKGVYAVEKTFETDGVYHVIAHTNARE-MHVMPEVKVAV 84
GK0969(45-128)                       GDGRYSVEKTFTEAGTYSVVAHVTARD-MHNMPKKDIVA 84
BL05305(45-129)                      GKGVYSMEKAFPEDGHYKVQVHVTAKK-QHTMPVADIKV 85
BSU30660(44-127)                     -KGVYRLETTFKEDGVYTVQSHVTAKK-QHSMPTLKVQV 84
OB2488(50-134)                       ENGTYTASYTFEEEKVYEMYAHTTAEA-IHSMPFKTVIV 85
B14911_05359_1(53-137)               GKGIYQAAKTFGKDGDYIVQVHVTARD-MHTMPKAEVQA 85
BH0678_1(45-129)                     GGGVYRVTYEFPEDGLYFVQPHVTARD-MHRMPLYELTI 85
ABC0230(45-129)                      ENGVYEITHTFDEDGIYIVQTHVTARD-MHVMPKQMIVA 85
ABC4088(44-127)                      ESGLYQAPLALEEAGIYMVQVHVTARG-MHVMPTQPLFA 84
BH0983(47-131)                       EDGIYLVEYEFPEDGIYFVQAHVTARG-LHVMPTERLIV 85
B14911_09907(34-118)                 GEGVYQLTKKFEQDGVYIIKVHASSGG-SLIMPQKQFVV 85
ExigDRAFT_1796(51-135)               KPGVYSIEKKLPKEGLYYIKVHASSNG-AMIMPTRQFIV 85
BAA83944_1(46-130)                   QDGAFNLPLTFESDGLYYMKVDVTHNG-DTIMPTAQLIV 85
BH1853(46-130)                       QDGAFNLPLTFESDGLYYMKVDVTHNG-DTIMPTAQLIV 85
OB3282(48-131)                       -NGEYSVETSFDQDGIYHMKVTASKNN-ATIMPTKQFIV 84
BCE_G9241_4093_2(163-245)            NKGEYETKHTFKENGAYKVKVHVRKGE-LHEHKEETIEV 83
BC4088_2(165-247)                    NKGEYETKHTFKENGAYKVKVHVRKGE-LHEHKEETIEV 83
RBTH_02670_2(165-247)                NKGEYETKHTFKENGAYKVKVHVRKGE-LHEHKEETIEV 83
BcerKBAB4DRAFT_2040_2(158-240)       NKGEYESKHTFKENGAYKVKVHVRKGE-LHEHKEETVEV 83
BT9727_3829_2(163-245)               NKGEYETKHTFKENGSYKVKVHVKKGE-LHEHKEETVEV 83
Bant_01004966_2(169-251)             NKGEYETKHTFKENGSYKVKVHVKKGE-LHEHKEETVEV 83
BA4310_2(163-245)                    NKGEYETKHTFKENGSYKVKVHVKKGE-LHEHKEETVEV 83
BCE4157_2(163-245)                   NKGEYETKHTFKENGSYKVKVHVKKGE-LHEHKEEXVEV 83
BCZK3845_2(163-245)                  NKGEYETKYTFKEKGSYKVKVHVKKGE-LHEHKEETVEV 83
Bcer98DRAFT_3614(94-176)             NKGEYVGKYTFKESGKYKVKVHVRKGD-LHEHKEETVEV 83
B14911_05359_2(187-271)              GEGKYSAEHSFAEAGSYTVTVHVENSEGLHEHSDFPLTV 85
BH0678_2(159-242)                    DVGQYEVSHTFADAGEYHVVFHIEDDTGLHEHIHEALIV 84
ExigDRAFT_0574(52-137)               KTGTYQAEAKLA-EGEYEGLYHINDKNGLHHMDKISFVV 86
                                       * :     :     *                 .
consensus/80%                        tpG.YtsphoFtpsG.YhlhsHspttp.hH.h.p.pl.V
```

BA4310 is homologous to proteins GBAA4310 from *B. anthracis* str. "Ames Ancestor," BAS3998 from *B. anthracis* str. Sterne, and BT9727_ 3829 from *B. thuringiensis* serovar konkukian str. 97-27.

Figure 4.1g: Multiple sequence alignment of 104 amino acid residue NTGFIG domain.

```
Secondary structure                    EEEEE            HH   H               HHHHH
BCZK2413_2(120-222)                VYNTGFIGVVFADLCSIDRFNFEF---EMGMLTKLMKDMIIPVKELFLR
BT9727_2444_2(120-222)             VYNTGFIGVVFADLCSIDRFNFEF---EMGMLTKLMKDMIIPVKELFLR
BA2665_2(120-222)                  VYNTGFIGVVFADLCSIDRFNFEF---EMGMLTKLMKDMIIPVKELFLR
Bant_01003317_2(124-226)           VYNTGFIGVVFADLCSIDRFNFEF---EMGMLTKLMKDMIIPVKELFLR
BCE2700_2(122-224)                 VFNTGFIGVVFADLCSIDRFNFEF---EMGMLTKLMKDMIIPVKELFLR
BCE_G9241_CNI_0263_2(122-224)      VFNTGFIGVVFADLCSIDRFNFEF---EMGMLTKLMKDMIIPVKELFLR
BcerKBAB4DRAFT_0535_2(120-222)     VFNTGFIGVVFADLSSIDRFNFEF---EMGMLTKLMKDMIIPVKELFLR
BC2674_2(122-224)                  VFNTGFIGVVFADLSSIDRFNFEF---EMGMLTKLMKDMIIPVKELFLR
Bcer98DRAFT_0128_2(122-224)        VFNTGFIGVVFADLSSIDRFNFEF---EMNMLFKLMKDMIIPVKELFLR
BA2665_1(16-119)                   ISNTGFIGSVFIDTLELQKKSYYFARKKLQIVHHVLDGLSGATSSLFKE
Bant_01003317_1(20-123)            ISNTGFIGSVFIDTLELQKKSYYFARKKLQIVHHVLDGLSGATSSLFKE
BCZK2413_1(16-119)                 ISNTGFIGSVFIDTLELQKKSYYFARKKLQIVHHVLDGLSGATSSLFKE
BT9727_2444_1(16-119)              ISNTGFIGSVFIDTLELQKKSYYFARKKLQIVHHVLDGLSGATSSLFKE
BcerKBAB4DRAFT_0535_1(16-119)      ISNTGFIGSVFIDTLELQKKSYYFARKKLQIVHHVLDGLSGATSALFKE
BCE2700_1(16-121)                  ISNTGFIGSVFIDTLELQKKSYYFARKKLQIVHHVLDGLSGATSSLFKE
BCE_G9241_CNI_0263_1(16-121)       ISNTGFIGSVFIDTLELQKKSYYFARKKLQIVHHVLDGLSGATSSLFKE
BC2674_1(16-121)                   ISNTGFIGSVFIDTLELQKKSYYFARKKLQIVHHVLDGLSGATSALFKE
Bcer98DRAFT_0128_1(16-121)         ISNTGFIGSVFIDTLELQKKSYYFSRKKLQIVHHVLDGLAEATSSLFHE
                                   : ****** ** *  .::: .: *   :: :: :::..:  ... ** .
consensus/80%                      l.NTGFIGsVFhDhhplp+hsa.F...chthlp+lhcsh.hssppLFhc


Secondary structure                                  EEEE          HHHHHHHHHHH
BCZK2413_2(120-222)                HNVPAYISTSHLEEQNKLGFVLSIKPYDERAEADLYFEAYLKERGL
BT9727_2444_2(120-222)             HNVPAYISTSHLEEQNKLGFVLSIKPYDERAEADLYFEAYLKERGL
BA2665_2(120-222)                  HNVPAYISTSHLEEQNKLGFVLSIKPYDERAEADLYFEAYLKERGL
Bant_01003317_2(124-226)           HNVPAYISTSHLEEXNKLGFVLSIKXYDERAEADLYFEAYLKERGL
BCE2700_2(122-224)                 HNVPAYISTSHLEEQNKLGFVLSVKPYDERAEADLYFEAYLKERGL
BCE_G9241_CNI_0263_2(122-224)      HNVPAYISTSHLEEQNKLGFVLSVKPYDERAEADLYFEAYLKERGL
BcerKBAB4DRAFT_0535_2(120-222)     HNVPAYISTSHLEEQNKLGFVLSVKPYDERAEADLYFETYLKERGL
BC2674_2(122-224)                  HNVPAYISTSHLEEQNKLGFVLSVKPYDERAEADLYFEAYLKERGL
Bcer98DRAFT_0128_2(122-224)        HNIPAYISTSHLETQNKVGFVLSIKPYDERAEADLYFETYLKERGL
BA2665_1(16-119)                   HNISAYMSCVYLHKQKKIGFVLSTKPFEQ-SDGVAYFINYLIEKNF
Bant_01003317_1(20-123)            HNISAYMSCVYLHKQKKIGFVLSTKPFEQ-SDGVAYFINYLIEKNF
BCZK2413_1(16-119)                 HNISAYMSCVYLHKQKKIGFVLSTKPFEQ-SDGVAYFINYLIEKNF
BT9727_2444_1(16-119)              HNISAYMSCVYLHKQKKIGFVLSTKPFEQ-SDGVAYFINYLIEKNF
BcerKBAB4DRAFT_0535_1(16-119)      HNTAAYMSCVYLHKQKKIGFVLSTKPFEQ-SDGVSYFINYLIEKNF
BCE2700_1(16-121)                  HNISAYMSCVYLHKQKKIGFVLSTKPFEQ-SDGVAYFVNYLIEKNF
BCE_G9241_CNI_0263_1(16-121)       HNISAYMSCVYLHKQKKIGFVLSTKPFEQ-SDGVAYFVNYLIEKNF
BC2674_1(16-121)                   HNISAYMSCVYLHKQKKIGFVLSTKPFEQ-SDGVAYFVNYLIEKNF
Bcer98DRAFT_0128_1(16-121)         HEVAAYISCVYLHKQKKIGFVLSTKLFEQ-TDGIAYFKNYLIEKNF
                                   *: .**:*  :*.  :*:***** * ::: ::.  **  ** *:.:
consensus/80%                      HNlsAYhSssaLccQpKlGFVLShKPa-p.u-ushYF.sYLhE+sh
```

BA2665 is homologous to proteins GBAA2665 from *B. anthracis* str. "Ames Ancestor," BAS2482 from *B. anthracis* str.Sterne. BT9727_2444 is homologous to protein BCZK2413 from *B. cereus* E33L.

Contd…..

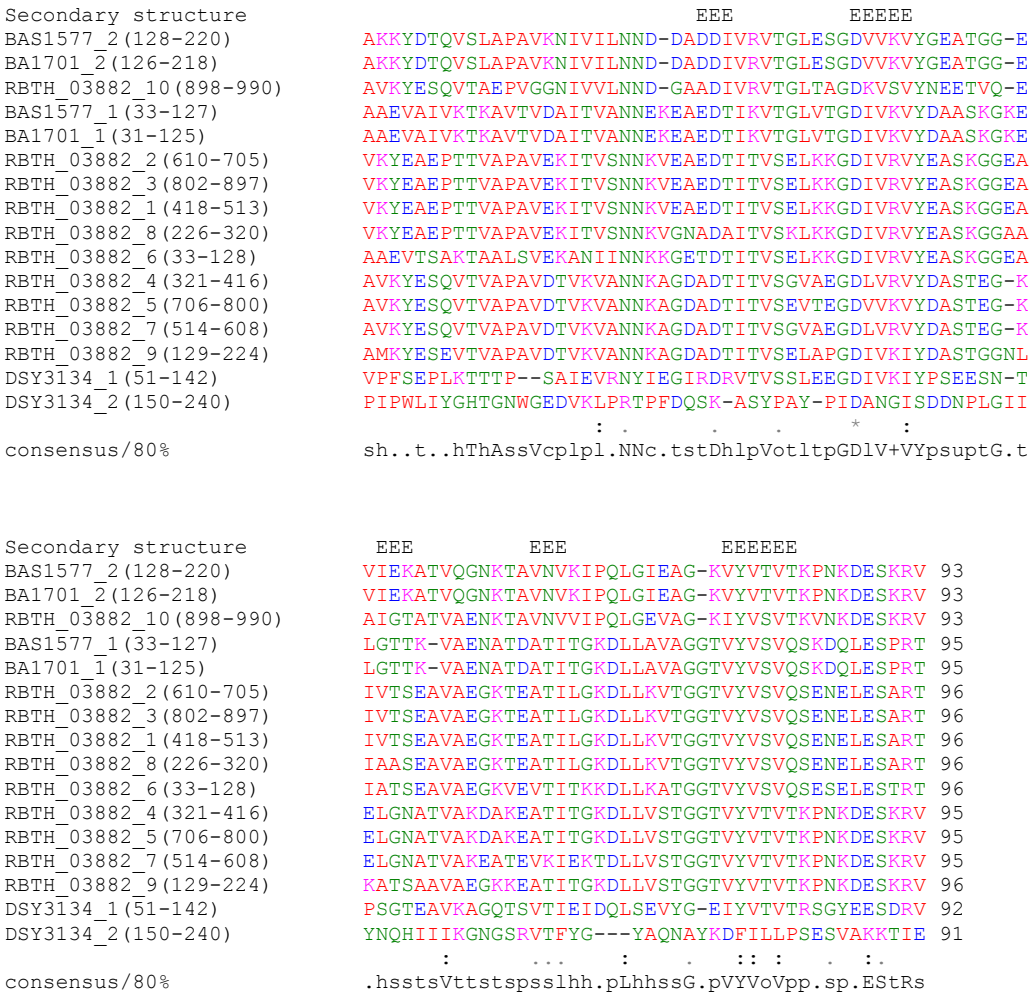
```
Secondary structure
BCZK2413_2(120-222)                 FIG-DEEDDIDK 103
BT9727_2444_2(120-222)              FIG-DEEDDIDK 103
BA2665_2(120-222)                   FIG-DEEDDIDK 103
Bant_01003317_2(124-226)            FIG-DEEDDIDK 103
BCE2700_2(122-224)                  FIG-DEEDDIDK 103
BCE_G9241_CNI_0263_2(122-224)       FIG-DEEDDIDK 103
BcerKBAB4DRAFT_0535_2(120-222)      FIG-DEEDDIDK 103
BC2674_2(122-224)                   FIG-DEEDDIDK 103
Bcer98DRAFT_0128_2(122-224)         FIG-DEEDEMDK 103
BA2665_1(16-119)                    YG--NEEVEYQE 104
Bant_01003317_1(20-123)             YG--NEEVEYQE 104
BCZK2413_1(16-119)                  YG--NEEVEYQE 104
BT9727_2444_1(16-119)               YG--NEEVEYQE 104
BcerKBAB4DRAFT_0535_1(16-119)       YG--NEEVEYQE 104
BCE2700_1(16-121)                   YGNHDEDVEYQE 106
BCE_G9241_CNI_0263_1(16-121)        YGNHDEDVEYQE 106
BC2674_1(16-121)                    YGGHDEDVEYQE 106
Bcer98DRAFT_0128_1(16-121)          YGKTDQEVEYQE 106
                                    :   ::: : ::
consensus/80%                       ah..sEEs-hpc
```

Figure 4.1h: Multiple sequence alignment of 36 amino acid residue NxGK repeat.

```
Secondary structure                                  HHHHHHHH
BT9727_3378_2(139-176)             YKTTISAQFEYNRFTRDFFEDPNNKGKSKADAIAAWNE 38
BcerKBAB4DRAFT_0944_2(139-176)     YKTTISAQFEYNRFTRDFFEDPNNKGKSKADAIAAWNE 38
BA3686_2(139-176)                  YKTTISAQFEYNRFTRDFFEDPNNKGKSKADAIAAWNE 38
BCZK3328_2(139-176)                YKTTISAQFEYNRFTRDFFEDPNNKGKSKADAIAAWNE 38
BCE_G9241_3579_2(139-176)          YKTTISSQFEYNRFTRDFFEDPNNKGKSKADAIAAWNE 38
BCE3645_2(139-176)                 YKTTISPQFEYNRFTRDFFEDPNNKGKTKADVIAAWNE 38
RBTH_03615_2(139-176)              YKTTIGAQFEYNRFTRDFFEDPNNKGKAKADAIAAWNE 38
BC3626_2(139-176)                  YKTTIGTQFEYNRFTRDFFEDPNNKGKAKADAIAAWNE 38
B14911_25780_2(138-175)            YKSEIGRQFEYNQFIRDYYADQKNQGKSRAEAIAAWML 38
RBTH_03615_1(94-129)               FKEKIGTNFRFTVALQKFFK--ENVGKTYEDAVAFWHE 36
BC3626_1(94-129)                   FKEKIGTNFRFTVALQKFFK--ENVGKTYEDAIAFWHE 36
BT9727_3378_1(94-129)              FKEKIGANFRFTVALQKFFK--ENIGKTYEDAVAFWHE 36
BCZK3328_1(94-129)                 FKEKIGANFRFTVALQKFFK--ENIGKTYEDAVAFWHE 36
BA3686_1(94-129)                   FKEKIGANFRFTVALQKFFK--ENIGKTYEDAVAFWHE 36
BCE_G9241_3579_1(94-129)           FKEKIGANFRFTVALQKFFK--ENVGKTYEDAITFWYE 36
BCE3645_1(94-129)                  FKEKIGANFRFTVALQKFFK--ENVGKTYEDAITFWYE 36
BcerKBAB4DRAFT_0944_1(94-129)      FKEKIGANFRFTVALQKFFK--ENVGKTYEDAITFWYE 36
B14911_25780_1(93-128)             FKSVIGSHFHFSTYIQDYFK--HNPGKTYNDAVSAWHE 36
                                   :*  *. :*.:.   :.::   .* **:  :.:: *
consensus/80%                      aKppIuspFcashhhpcFFc..pNhGKohtDAluhW.E
```

BT9727_3378 is homologous to protein BCZK3328 from *B. cereus* E33L. BA3686 is homologous to proteins GBAA3686 from *B. anthracis* str. "Ames Ancestor," BAS3417 from *B. anthracis* str. Sterne, and Bant_ 01004341 from *B. anthracis* str. A2012.

Figure 4.1i: Multiple sequence alignment of 95 amino acid residue VYV domain.



BA1701 is homologous to proteins GBAA1701 from *B. anthracis* str. "Ames Ancestor," and Bant_01002313 from *B. anthracis* str. A2012.

Figure 4.1j: Multiple sequence alignment of 75 amino acid residue KEWE domain.

```
Secondary structure            HHHHHHHHHHHHHHHHH   HH
RBTH_06405_4(259-331)         REKALDALQWTIEEKEKLTDNQLLQQYTMQWLKNHRLWTPVVRYWNGSPY
pBMB165_3(175-247)            REKALEALQWTIEEKEKLIDNQLLQQYTMKWLKRHRLWTPVVRYWNGSPY
pE33L466_0092_4(259-328)      KRKALEALRWTIEEKEKLDEKQLLKVFNQKWLIKQKLWTPLKRYWKGSPY
RBTH_06405_3(109-183)         KEKALQLLKWLIEEEEKLPPQKLLQIYGQKWLIEHRLSAPLRVIWNGSPY
pBMB165_2(25-99)              KEKALQLLKWIIEEEEKVSPQKLLQIYGQKWLNERRLSAPLRVIWDGSPY
BA3147_2(109-183)             KEKALEALKWTVEEKEKLSKVELLKFYSKKWLEKNKLSAPLVMYWNGSPY
Bant_01003795_1(25-99)        KEKALEALKWTVEEKEKLSKVELLKFYSKKWLEKNKLSAPLVMYWNGSPY
BAS2924_2(116-190)            KEKALEALKWTVEEKEKLSKVELLKFYSKKWLEKNKLSAPLVMYWNGSPY
BAS2924_3(191-265)            KEKALEALKWTVEEKEKLSKVELLKFYSKKWLEKNKLSAPLVMYWNGSPY
pE33L466_0092_2(109-183)      KEKALTILKWIIEEKEGLSQEKLLELYGKKWLEKNKLGAPLAMYWNSSPY
RBTH_06405_2(184-258)         KDKTLQALKWTIEKKEKLNVDQLKNIYDNKWLVQSGLSGACQLYWNDSPY
pBMB165_1(100-174)            KEKALQALKWTIEEKEKLNPDQLKNIYENKWLTQLGLRGACQLYWNDSPY
BAS2924_4(266-340)            KEKALVALRWTIEEKEKLTSFQLLQVYSVKWLTIHNLISPCQIFWNNSPY
Bant_01003795_2(100-174)      KEKALVALRWTIEEKEKLTSFQLLQVYSVKWLTIHNLISPCQIFWNNSPY
BA3147_3(184-258)             KEKALVALRWTIEEKEKLTSFQLLQVYSVKWLTIHNLISPCQIFWNNSPY
pE33L466_0092_3(184-258)      KEKALEALKWTIEEKEGLTPKQLLDVYNIKWLQTHRLASACQIIWGNSPF
BA3147_1(34-108)              RELSKRVTKYLIETILKWNEEDIKQKWNTPLIIKYRLLGALKHGYDNSPY
BAS2924_1(41-115)             RELSKRVTKYLIETILKWNEEDIKQKWNTPLIIKYRLLGALKHGYDNSPY
RBTH_06405_1(34-108)          NQLARRVTKYLVTKILNWNEEEIKQNWNNKLIAKYRLRGVLKHKYNNSPY
pE33L466_0092_1(34-108)       NKMARRVLTYLLNSILKWNKEDIRKKWNTKLLVKYRLRGLLKHRYENSPF
                              .  :     : :          .: . :    :    *       : .**:
consensus/80%                 +EKALpsL+WhlEccEKls..pLhphas.KWL.p.pL.ush.hhWssSPY
```

```
Secondary structure           HHHHHH
RBTH_06405_4(259-331)         AMINDLYPNKYIKSSFSGYINKF-- 73
pBMB165_3(175-247)            AMINDLYPNKYLKSSFRGYINKS-- 73
pE33L466_0092_4(259-328)      EMLIALYPNRFSKNMLKGYM----- 70
RBTH_06405_3(109-183)         AMINDLYPNRFKEWEFNKAPNKFWT 75
pBMB165_2(25-99)              AMINDLYPNRFKEWEFTKAPNKFWT 75
BA3147_2(109-183)             AMINSLYPNKFKEWEFSMTPNNFWT 75
Bant_01003795_1(25-99)        AMINSLYPNKFKEWEFSMTPNNFWT 75
BAS2924_2(116-190)            AMINSLYPNKFKEWEFSMTPNNFWT 75
BAS2924_3(191-265)            AMINSLYPNKFKEWEFSMTPNNFWT 75
pE33L466_0092_2(109-183)      AMINDLYPRRFKEWEFGMTPNNFWT 75
RBTH_06405_2(184-258)         AMINDLYPGQFKEWEFKMTPNGFWT 75
pBMB165_1(100-174)            AMINDLYPNQFKEWEFKMTPSGFWT 75
BAS2924_4(266-340)            SMINELYPGQNKEWEYKFTPTGFWT 75
Bant_01003795_2(100-174)      SMINELYPGQNKEWEYKFTPTGFWT 75
BA3147_3(184-258)             SMINELYPGQNKEWEYKFTPTGFWT 75
pE33L466_0092_3(184-258)      RMINDLYIDRFKEWEFRVTPVGYWS 75
BA3147_1(34-108)              KMIEDLYPNRFKEWEFGMAPLNFWT 75
BAS2924_1(41-115)             KMIEDLYPNRFKEWEFGMAPLNFWT 75
RBTH_06405_1(34-108)          AMINDLYPNQFKEWEFRMTPLNFWT 75
pE33L466_0092_1(34-108)       KAINDLYPNQFKEWEFGMTPLNFWT 75
                                :  **  :  :
consensus/80%                 tMINsLYPspaKEWEFphsP.tFWT
```

BA3147 is homologous to protein GBAA3147 from *B. anthracis* str. “Ames Ancestor.”

Figure 4.1k: Multiple sequence alignment of 59 amino acid residue AFL domain.

```
Secondary structure                          EEE                    HHHHHH
BAS2851_1(20-78)                    LEYQQSRFYVTRIPKDFLSIARKRFSIPTDDQIIAFLSCNL
BA3065_1(13-71)                     LEYQQSRFYVTRIPKDFLSIARKRFSIPTDDQIIAFLSCNL
Bant_01003715_1(16-74)              LEYQQSRFYVTRIPKDFLSIARKRFSIPTDDQIIAFLSCNL
BcerKBAB4DRAFT_1832_1(14-72)        LEYQQSRFYVTRIPKDFLSVAKKRFSIPIDDRIFAFLSCNL
RBTH_02124_1(13-71)                 LEFQQSRFYVTRIPKDFLSIAQKRFSIPTEDQIVAFLSCNL
Bant_01003715_2(164-225)            LEPDNGLFVETHISDKKLKAIEVRFIIPIEEQIIAFLDTSV
BAS2851_2(168-229)                  LEPDNGLFVETHISDKKLKAIEVRFIIPIEEQIIAFLDTSV
BA3065_2(161-222)                   LEPDNGLFVETHISDKKLKAIEVRFIIPIEEQIIAFLDTSV
BcerKBAB4DRAFT_1832_2(162-223)      LEPDNGLFVDTHISHKKLKEIGAKYIIPKEEKIIAFLDTSV
                                    ** ::. *  *:*... *.    :: ** :::*.***. .:
consensus/80%                       LE.ppuhFh.T+IscchLphhphRF.IPh--pIlAFLsssl

Secondary structure                        EEE
BAS2851_1(20-78)                    FG---SGKYGVYFTSSGLYWK 59
BA3065_1(13-71)                     FG---SGKYGVYFTSSGLYWK 59
Bant_01003715_1(16-74)              FG---SGKYGVYFTSSGLYWK 59
BcerKBAB4DRAFT_1832_1(14-72)        FG---SGKYGVYFTSSGLYWK 59
RBTH_02124_1(13-71)                 LG---SGKYGVYFTSSGLYWK 59
Bant_01003715_2(164-225)            LGNMGKGSDGVLICQSGIYFR 62
BAS2851_2(168-229)                  LGNMGKGSDGVLICQSGIYFR 62
BA3065_2(161-222)                   LGNMGKGSDGVLICQSGIYFR 62
BcerKBAB4DRAFT_1832_2(162-223)      LGNLGKGSDGVLICEPGIYFR 62
                                    :*   .*. ** : ..*:*::
consensus/80%                       hG...pGp.GVhhspSGlYa+
```

BA3065 is homologous to protein GBAA3065 from *B. anthracis* str. “Ames Ancestor.”

Figure 4.1l: Multiple sequence alignment of 53 amino acid residue RIDVK repeat.

```
Secondary structure     EEEE      EEE                EEEEE
BA0482_1(4-56)          IEIHTQGGLKHKVQTEVYNAEALNTKLNDNDLITVLIGDFIIQRIDVKRIIPL 53
BA0482_2(67-119)        VEVHTNAGKVIEITTNDYDPIYLNEQLNNNNTITVVIGDYIFSRIDVKQVVPV 53
                        :*:**:.*   :: *: *:.  ** :**:*: ***:***:*:.*****:::*:
consensus/80%           lElHTpuGhhhclpTpsYss.hLNppLNsNshITVlIGDaIhpRIDVKpllPl
```

BA0482 is homologous to proteins GBAA0482 from *B. anthracis* str. “Ames Ancestor,” BAS0458 from *B. anthracis* str. Sterne, and Bant_ 01001108 from *B. anthracis* str. A2012.

Figure 4.1m: Multiple sequence alignment of a) 41 amino acid residue AGQF repeat and b) 42 amino acid residue GSAL repeat.

```
(a)   Secondary structure                    HHHH
      BA4081_1(10-50)          SIGMYLSELQKGTESSRLLAESMAKEIDGKMKIDLGPAGQF 41
      BA4081_2(172-212)        NIQTLINGMQIGALSLPQVAQTMGLDIKSNVQVDLGEAGQF 41
                               .*   :. :* *: *   :*::*. :*..::::*** **** 
      consensus/80%            sIthhlsthQhGs.S...lApoMuh-IcuphplDLG.AGQF


(b)   Secondary structure              EEE         HHH
      BA4081_1(292-333)        GSKSGSELGQGIISQDGYIKGSALQVVGSAHNAFSTINGSPA 42
      BA4081_1(334-375)        GNQGGQGFGSGIVNQKGYIRGSALEAVTPAHTGFNTINGTPQ 42
                               *.:.*. :*.**:.*.***:****:.* .**..*.****:* 
      consensus/80%            GspuGpthGpGIlsQcGYI+GSALpsVssAHsuFsTINGoPt
```

BA4081 is homologous to proteins GBAA4081 from *B. anthracis* str. "Ames Ancestor," BAS3792 from *B. anthracis* str. Sterne, and Bant_01004731 from *B. anthracis* str. A2012.

Figure 4.2a: PxV-57 domain.

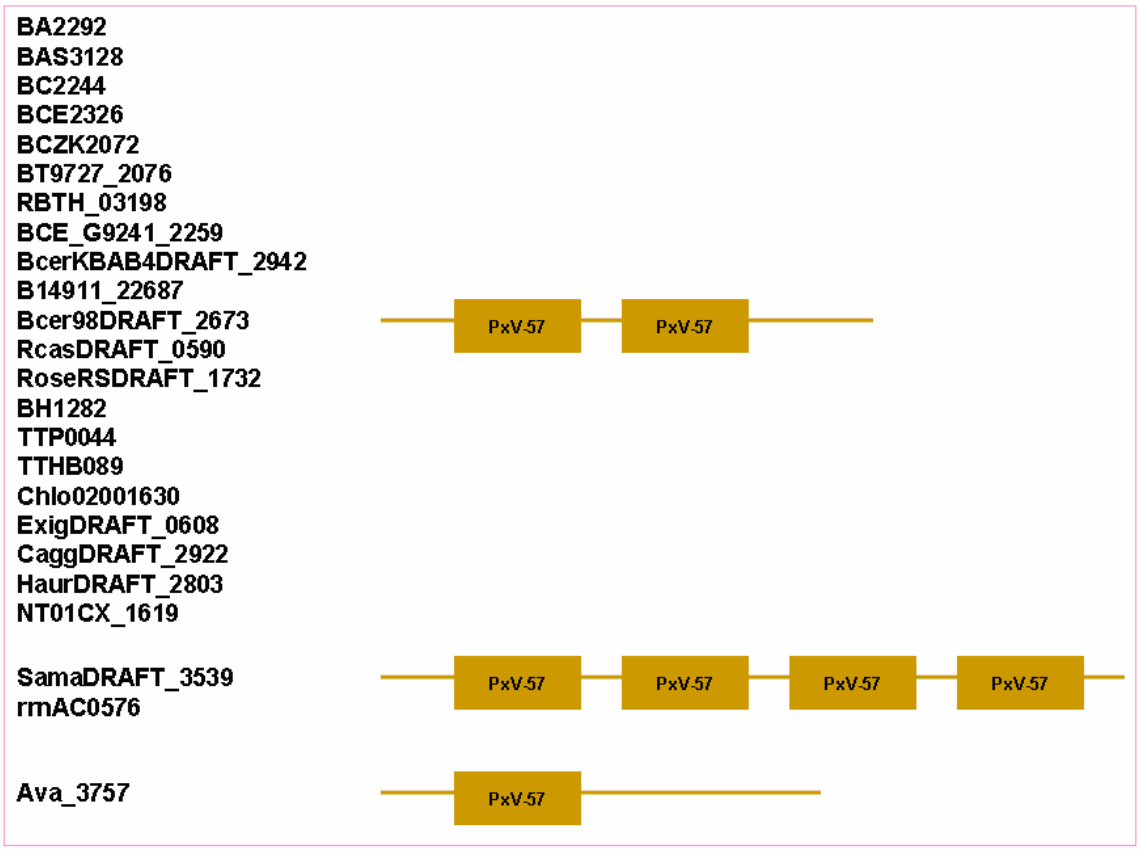


Figure 4.2b: FxF-122 domain.

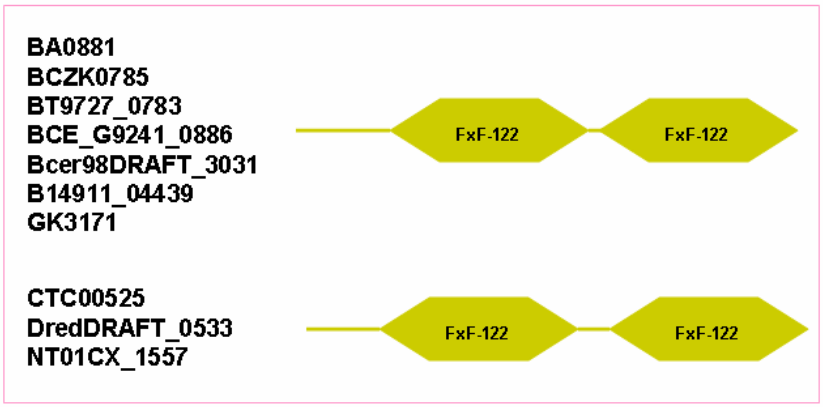


Figure 4.2c: YEFF-111 domain.

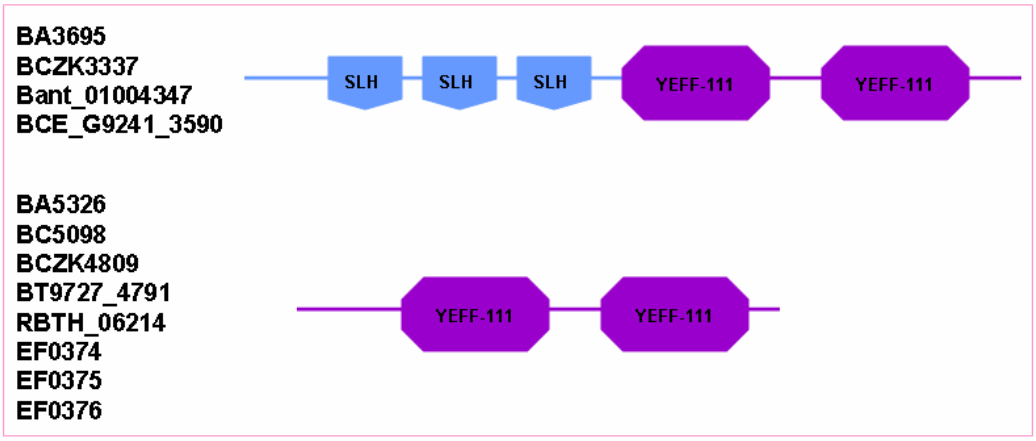


Figure 4.2d: IMxxH-109 domain.

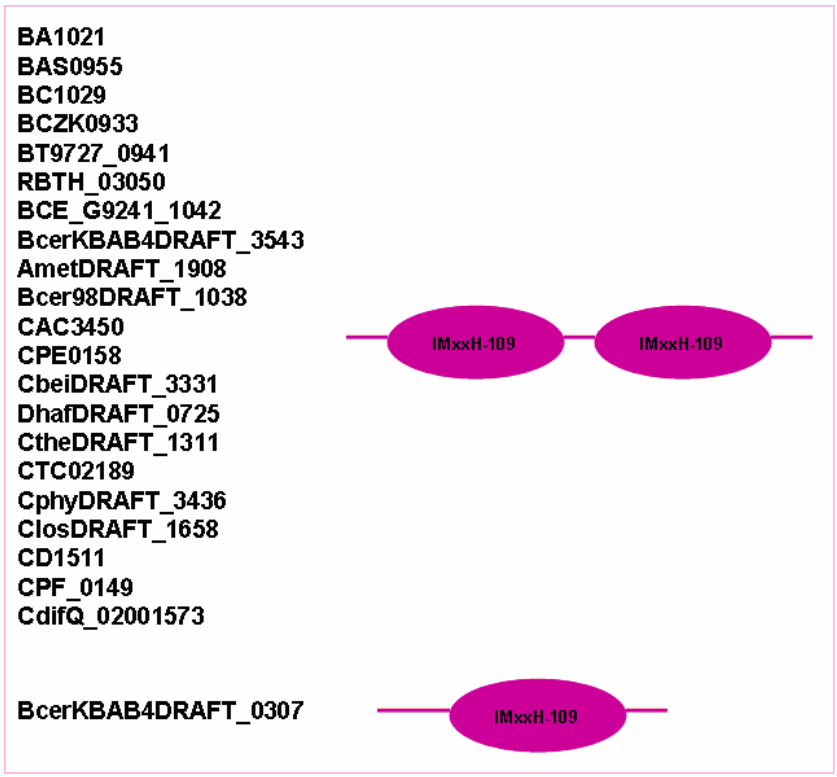


Figure 4.2e: VxxT-103 domain.

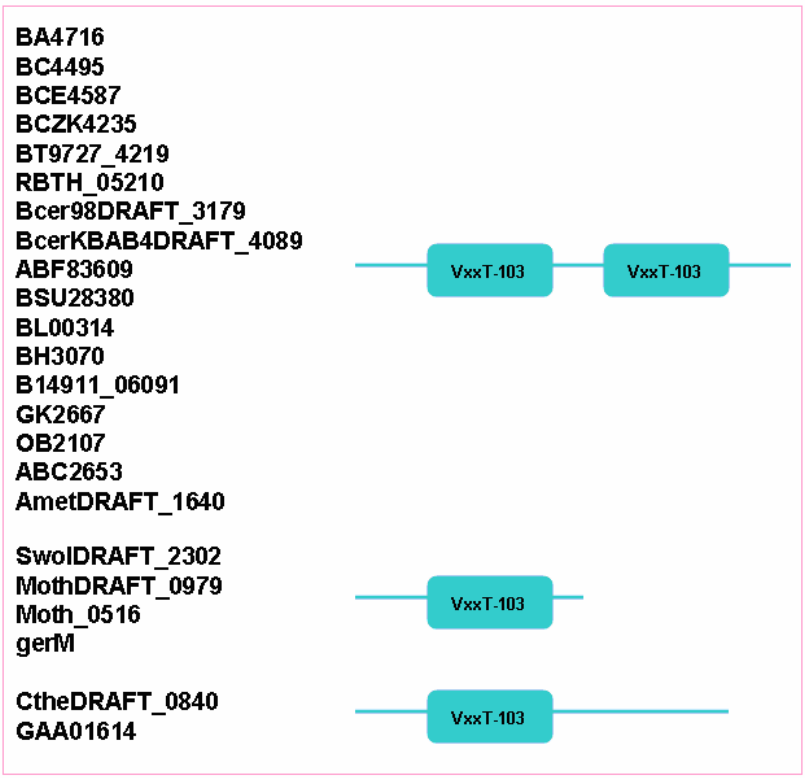


Figure 4.2f: ExW-84 domain.

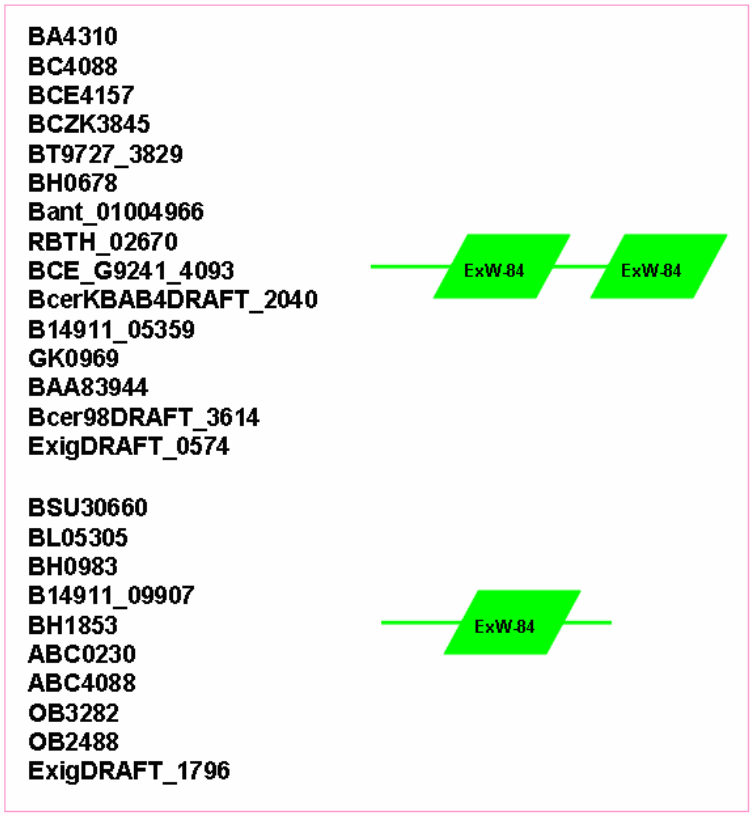


Figure 4.2g: NTGFIG-104 domain.

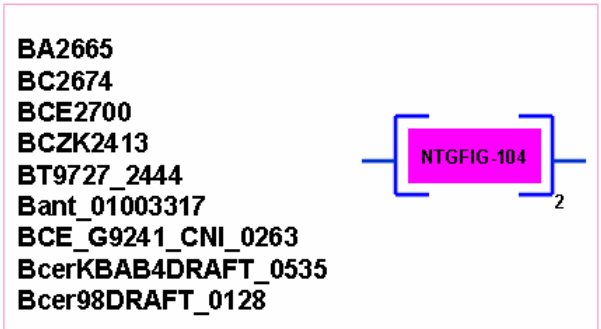


Figure 4.2h: NxGK-36 repeat.

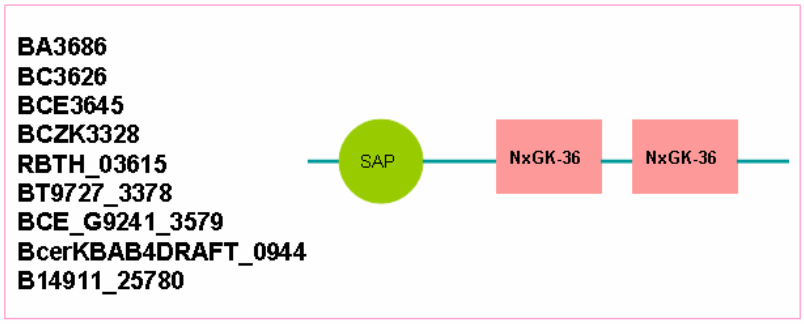


Figure 4.2i: VYV-95 domain.

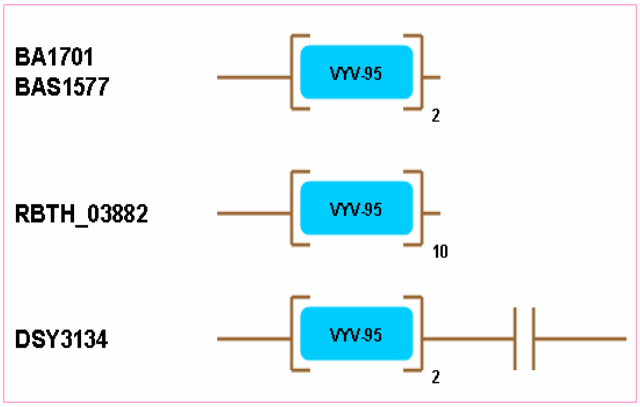


Figure 4.2j: KEWE-75 domain.

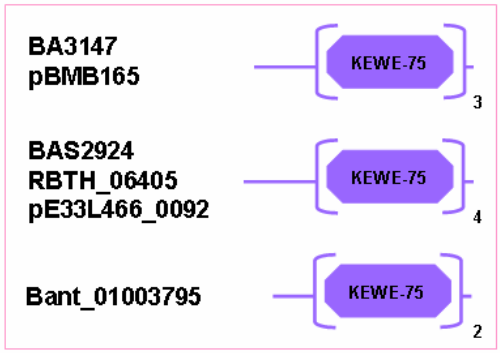


Figure 4.2k: AFL-59 domain.

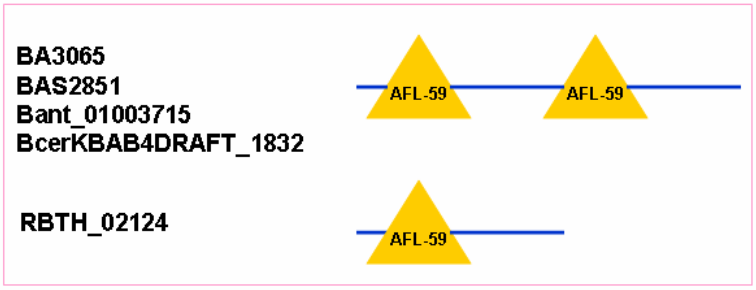


Figure 4.2l: RIDVK-53 repeat.

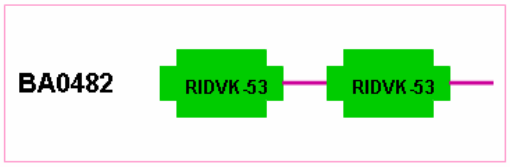
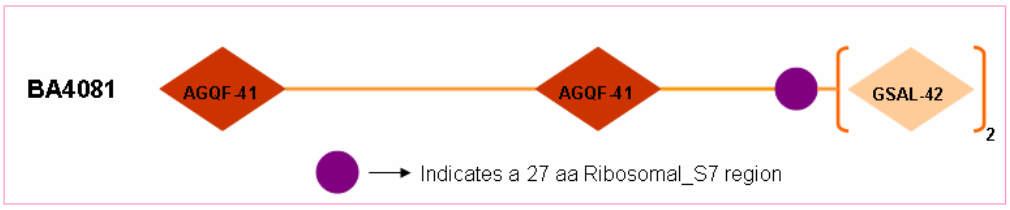


Figure 4.2m: AGQF-41 & GSAL-42 repeat.



The domain architechture diagrams of the representative repeats and domains from various proteins along with their GENE or SWall identifiers. (a) PxV domain, (b) FxF domain, (c) YEFF domain, (d) IMxxH domain, (e) VxxT domain, (f) ExW domain, (g) NTGFIG domain, (h) NxGK repeat (i) VYV domain, (j) KEWE domain, (k) AFL domain, (l) RIDVK repeat, (m) AGQF repeat and GSAL repeats.

4.4 Conclusions

1. A systematic analysis using computational tools identified 4 novel repeats and 10 domains corresponding to the *B. anthracis* str. Ames proteome.

2. The NxGK repeats are associated with SAP domain. The SAP domain is a DNA-binding motif that is involved in chromosomal organization. Therefore, we believe that these repeats also participate in a similar function.

3. The YEFF domain containing proteins are associated with RGD motif and may be involved in cell adhesion.

4. From the presence of VYV and AFL domains in all the *B. anthracis* species and their absence in *B. cereus* genomes, we identified some differences in these two genomes that are otherwise closely related.

5. The identification of novel repeats and domains corresponding to *B. anthracis* str. Ames proteome may be useful for annotation.

4.5 References

Aravind, L. & Koonin, E. V. (2000). SAP-a putative DNA-binding motif involved in chromosomal organization. *Trends Biochem. Sci.* **25,** 112-114.

Akula, S. M., Pramod, N. P., Wang, F. Z. & Chandran, B. (2002). Integrin α1β2 (CD 49c/29) Is a Cellular Receptor for Kaposi's Sarcoma-Associated Herpesvirus (KSHV/HHV-8) Entry into the Target Cells. *Cell*, **108,** 407-419.

Dixon, T. C., Meselson, M., Guillemin, J. & Hannam, P. C. (1999). Anthrax. N. Engl. *J. Med.* **341,** 815-826.

D'Souza, S. E., Ginsberg, M. H. & Plow, E. F. (1991). Arginyl-glycyl-aspartic acid (RGD): a cell adhesion motif. *Trends Biochem. Sci.* **16,** 246-250.

Falquet, L ., Pagni, M., Bucher, P., Hulo, N., Sigrist, C. J., Hofmann, K. & Bairoch, A. (2002). The PROSITE database, its status in 2002. *Nucleic Acids Res.* **30,** 235-238.

Guttmann, D. M. & Ellar, D. J. (2000). Phenotypic and genotypic comparisons of 23 strains from the Bacillus cereus complex for a selection of known and putative *B. thuringiensis* virulence factors. *FEMS Microbiol. Lett.* **188,** 7-13.

Kobe, B. & Deisenhofer, J. (1994). The leucine-rich repeat: A versatile binding motif. *Trends Biochem. Sci.* **19,** 415-421.

Leppla, S. H. (1995). Anthrax toxins, p. 543-572. In Moss, J., Iglewski, B., Vaughn, M. & Tu A.T. (ed.), Bacterial toxins and virulence factors in disease. Marcel Dekker, New York, N.Y.

Lupas, A., Englehardt, H., Peters, J., Santarius, U., Volker, S. & Baumeister, W. (1994). Domain structure of the *Acetogenium kivui* surface layer revealed by electron crystallography and sequence analysis. *J. Bacteriol.* **176,** 1224-1233.

Lupas, A. (1996). A circular permutation event in the evolution of the SLH domain? *Mol. Microbiol.* **20,** 897-898.

Navarre, W. W. & Schneewind, O. (1999). Surface proteins of Gram-positive bacteria and mechanisms of their targeting to the cell wall envelope. *Microbiol. Mol. Biol. Rev.* **63,** 174-229.

Okinaka, R., Cloud, K., Hampton, O., Hoffmaster, A., Hill, K., Keim, P., Koehler, T. *et al.* (1999). Sequence, assembly and analysis of pXO1 and pXO2. *J. Appl. Microbiol.* **87,** 261-262.

Pallen, M. J., Lam, A. C., Antonio, M. & Dunbar, K. (2001). An embarrassment of sortases-a richness of substrates? *Trends Microbiol.* **9,** 97-102.

Patti, J. M., Allen, B. L., McGavin, M. J. & Hook, M. (1994). MSCRAMM-mediated adherence of microorganisms to host tissues. *Annu. Rev. Microbiol.* **48,** 585–617.

Read, T. D., Peterson, S. N., Tourasses, N., Baillie, L. W., Paulsen, I. T., Nelson, K. E. *et al.* (2003). The genome sequence of *Bacillus anthracis Ames* and comparison to closely related bacteria. *Nature,* **423,** 81-86.

Szklarczyk, R. & Heringa, J. (2004). TRUST: Tracking Repeats Using Significance and Transitivity. *Bioinformatics.* **00,** 1-7.

Uchida, I., Sekizaki, T., Hashimoto, K. & Terakado, N. (1985). Association of the encapsulation of Bacillus anthracis with a 60 megadalton plasmid. *J. Gen. Microbiol.* **131,** 363-367.

Uchida, I., Hashimoto, K. & Terakado, N. (1986). Virulence and immunogenicity in experimental animals of Bacillus anthracis strains harbouring or lacking 110 MDa and 60 MDa plasmids. *J. Gen. Microbiol.* **132,** 557-559.

Yu, X., Christiano, C., Chini, S., He, M., Mer, G. & Chen, J. (2003). The BRCT domain is a phospho-protein binding domain. *Science,* **302,** 639-642.

CHAPTER 5

Identification and Analysis of Novel Amino Acid Sequence Repeats and Domains in Representative Archaeal Proteomes

5.1 Introduction

Archaea is a major division of living organisms. Although there is still uncertainty in the exact phylogeny of the groups, archaea, eukaryota and bacteria are the fundamental classifications in the three-domain system (see Figure 5.1). Carl Woese and George Fox identified archaea in 1977 based on the separation from other prokaryotes and 16S rRNA phylogenetic tree. The two groups, archaea and eukaryota were originally named the archaebacteria and eubacteria, treated as kingdoms or subkingdoms, which Woese and Fox termed 'Urkingdoms' (Woese & Fox, 1977). Like bacteria, archaea are single cell organisms lacking nuclei and are therefore classified as prokaryota — known as Monera in the five kingdom taxonomy. They were initially discovered in extreme environments, but have since been found in all types of habitats and may contribute upto 20% of total biomass (DeLong & Pace 2001). A single organism from this domain has been called an "Archaean" (Valentine, 2007).

Individual archaeans range from 0.1 μm to over 15 μm in diameter and some form aggregates or filaments upto 200 μm in length. They occur in various shapes, such as spherical, rod-shape, spiral, lobed or rectangular. A species of flat, square archaean that lives in hypersaline pools has been discovered. Archaea have no murein in their cell walls (Burns *et al.*, 2004). Phylogenetic analysis of small-subunit rRNA sequences distinguishes two distinct archaeal sub-domains: the euryarchaeotes and the crenarchaeotes (Woese, 1993). The euryarchaeotes include methanogens, halophiles and sulfur-reducing thermophiles. Although methanogenesis is uniform in two of the three major methanogenic euryarchaeal lineages, variations occur within the methanomicrobiales lineage (Danson, 1993). The latest branching methogenic euryarchaeal lineage, methanomicrobiales, gave rise to the extreme halophiles and sulfate-reducing archaea.

The euryarchaeota is further divided into nine families. They are as follows 1. Archaeoglobales (*Archaeoglobus fulgidus* DSM 4304), 2. Halobacteriales (*Halobacterium salinarium* NRC-1), 3. Methanobacteriales (*Methanobacterium thermoautotrophicum* str. Delta H*)* 4. Methanococcales (*Methanocaldococcus jannaschii* DSM 2661*),* 5. Methanopyrales (*Methanopyrus kandleri* AV19), 6. Methanosarcinales *(Methanosarcina acetivorans* str. C2A), 7. Thermococcales (*Pyrococcus abyssii* GE5), 8. Thermoplasmales (*Thermoplasma acidophilum* DSM 1728*)* and 9. Thermoplasmatales (*Picrophilus torridus* DSM 9790).

The crenarchaeotes share a 16S rRNA signature with the euryarchaeotes within the archaeal domain. Crenarchaeotes are in many instances sulfur-dependent thermophiles and have initially been regarded as more homogenous than the euryarchaotes (Woese *et al*., 1990). However, isolation of small-subunit rRNA from the open environment and the discovery of *Crenarhaeum symbiosum* has led to the characterization of deeply divergent lineages of low-temperature crenarchaeota (DeLong, 1992; Fuhrman *et al.*, 1992). The crenarchaeota is further divided into three families. They are 1. Desulfurococcales – *(Aeropyrum pernix* K1*),* 2. Sulfolobales *(Sulfolobus tokodaii* str. 7*)* and 3. Thermoproteales (*Pyrobaculum aerophilum* str. IM2).

Nanoarchaeota is the newly identified domain and the distribution of the nanoarchaeota is so far unknown. *Nanoarchaeum equitans* belongs to archaea. The cells of *N. equitans* are spherical and only about 400 nm in diameter. They grow attached to the surface of a specific archaeal host, a new member of the genus Ignicoccus. Owing to their unusual single stranded rRNA sequence, members remained undetectable by commonly used ecological studies based on the polymerase chain reaction (Huber *et al*., 2002).

Archaea are distinguished from other organisms by three major criteria: 1. Their 16S rRNA sequences are different from those of eubacteria and eukaryotes, 2. Their cell walls consist of glycosylated proteins rather than peptidoglycan structure in eubacteria. The *Thermoplasma* differ somewhat from the other archaea: they have no cell wall and their cell membranes contain tetra ether lipids with mannose and glucose subunits (Gaasterland, 1999) and 3. Their membrane lipids are unique, consisting entirely of derivatives of an ether linked isoprenoid structure (Kates, 1992). Archaeal tRNA and rRNA genes harbor unique archaeal introns which are neither like eukaryotic introns nor like bacterial introns. The archaeal challenge to phylogeny has continued with each new release of a completely sequenced archaeal genome. Archaea are highly diverse in terms of their physiology, metabolism and ecology. Presently, very few molecular characteristics are known that are uniquely shared by either all archaea or the different main groups within archaea. The evolutionary relationships among different groups within the euryarchaeota branch are also not clearly understood (Gao & Gupta, 2007).

The complete and nearly complete sequencing of archaeal genomes will provide data to infer properties of proteins that must have been present in a common ancestor, as well as properties that may pinpoint the basis of divergence. Since many proteins in these genomes are identified from genome sequencing projects, they are hypothetical and yet to be characterized. Therefore, in order to further characterize these hypothetical proteins we have carried out a systematic identification and analysis of the novel amino acid sequence repeats of all the available representative archaeal proteomes using computational tools.

We have identified and analyzed 56 domains and 38 repeats in 13 archaeal proteomes according to the representative phylogeny. These repeats and domains have not been reported before in archaeal proteomes. They are as follows: 1. *Aeropyrum pernix* K1 (1 domain), 2. *Sulfolobus tokodaii* str. 7 (7

domains and 5 repeats), 3. *Pyrobaculum aerophilum* str. IM2 (5 domains and 4 repeats), 4. *Archaeoglobus fulgidus* DSM 4304 (7 domains and 4 repeats), 5. *Halobacterium salinarium* NRC-1 (8 domains and 1 repeat), 6. *Methanobacterium thermoautotrophicum* str. Delta H (4 domains and 2 repeats), 7. *Methanocaldococcus jannaschii* DSM 2661 (5 domains and 2 repeats), 8. *Methanopyrus kandleri* AV19 (2 domains), 9. *Methanosarcina acetivorans* str. C2A (8 domains and 13 repeats), 10. *Pyrococcus abyssi* GE5 (4 domains), 11. *Thermoplasma acidophilum* DSM 1728 (4 domains), 12. *Picrophilus torridus* DSM 9790 (6 repeats) and 13. *Nanoarchaeum equitans* Kin4-M (1 domain and 1 repeat). We discuss the presence of these novel repeats and domains in proteins from other proteomes and their predicted secondary structure.

5.2 Methods

Various methods used to carry out the repeat identification analysis has been discussed in detail in Chapter 3.

5.3 Results and Discussion

From the sequence analysis using TRUST program, we identified 56 domains and 38 repeats that have not been reported before and are therefore novel. The detailed classification of 13 archaeal organisms for which *in silico* repeat identification has been carried out is shown in Figure 5.2. Each representative organism's proteome as shown in the figure is downloaded and the repeat analysis was carried out using the computational tools.

The Table 5a lists the genomes studied in this work, the toal number of proteins and the number of repeats identified by TRUST are mentioned. The known repeats present in each proteome are indicated. Also, the number of novel repeats and domains are indicated in the table. Lists of the proteins containing these novel repeats and domains of each representative organism are shown in Tables 5b to 5n. These tables indicate the names of the novel repeats and domains, their length, predicted secondary structure and their order, number of proteins identified from PSI-BLAST and the taxonomy of the organisms in other archaeal and bacterial genomes. The secondary structural elements aligned well in the multiple sequence alignment. We discuss each of these novel repeats and domains below.

Crenarchaeota
(*Aeropyrum pernix* K1, *Sulfolobus tokodaii* str. 7, *Pyrobaculum aerophilum* str. IM2)

I. *Aeropyrum pernix* K1*:* *Aeropyrum pernix* K1 proteome comprises 1 domain.

1. 90 amino acid residue PxG domain: The protein corresponding to the GENE_ID APE_0620 with a length of 1950 amino acid residues and described as hypothetical protein comprises of 90 amino acid residue region as three copies. The multiple sequence alignment corresponding to this domain is associated with PxG motif. The pair-wise identity between sequences corresponding to the PxG domain varied between 16-27%. The consensus

secondary structure is predicted to comprise of 8 β strands. This domain is *A. pernix* K1 specific as it occurs only in this proteome.

The representative table corresponding to PxG domain is shown in Table 5b.

II. *Sulfolobus tokodaii* str. 7: The *Sulfolobus tokodaii* str. 7 proteome comprises of 5 repeats and 7 domains.

1. 46 amino acid residue EYL repeat: The protein corresponding to GENE_ID ST0710 comprising 132 amino acid residues consists of 46 amino acid residue region as two copies. The multiple sequence alignment suggests EYL motif. The predicted secondary structure comprises of 1 α helix and 3 β strands. This repeat is specific to *S. tokodaii* str. 7 proteome. The sequence homology shared between the EYL repeat is 80%.

2. 44 amino acid residue LVVV repeat: The protein corresponding to GENE_ID ST1162 comprising 178 amino acid residues consists of 44 amino acid residue region as two copies. Further PSI-BLAST searches with sequence corresponding to region (37-80) as query identified two copies in *S. tokodaii* str. 7 and *S. acidocaldarius* DSM 639 and one copy each in *M. sedula* DSM 5348 and *M. thermoacetica* ATCC 39073. The multiple sequence alignment identified LVVV as conserved sequence motif. The pair-wise sequence identities corresponding to the LVVV repeat varied between 27-88%. The length of the proteins varied between 69 to 178 amino acid residues. The predicted secondary structure comprises of 1 α helix and 2 β strands.

3. 30 amino acid residue LIN repeat: The protein corresponding to GENE_ID ST1883 comprising of 450 amino acid residues consists of 30 amino acid residue region as four copies. Further PSI-BLAST searches with sequence corresponding to the region (51-80) as query identified four copies in *S. tokodaii* str. 7, *S. acidocaldarius* DSM 639, *S. solfataricus* P2 and *M. sedula* DSM 5348. The multiple sequence alignment identified LIN as conserved sequence motif. The pair-wise sequence identities corresponding to the LIN

repeat varied between 63-83%. The length of the proteins varied between 357 to 450 amino acid residues. The predicted secondary structure comprises of 4 α helices.

4. 43 amino acid residue KxK repeat: The protein corresponding to GENE_ID ST2173 comprising of 269 amino acid residues consists of 43 amino acid residue region as two copies. The predicted secondary structure comprises of 2 α helices and 2 β strands. The sequence homology shared between the KxK repeat is 25%. This repeat is specific to *S. tokodaii* str.7 proteome.

5. 48 amino acid residue GTY repeat: The protein corresponding to GENE_ID ST2253 comprising of 781 amino acid residues consists of 48 amino acid residue region as one copy. Further PSI-BLAST searches corresponding to the region (143-190) as query identified one copy in *S. tokodaii* str. 7, *S. acidocaldarius* DSM 639, *S. solfataricus* P2, *M. sedula* DSM 5348 and three copies in *P. torridus* DSM 9790. The multiple sequence alignment identified GTY as conserved sequence motif. The pair-wise sequence identities corresponding to the GTY repeat varied between 27-79%. The length of the proteins varied between 337 to 781 amino acid residues. The predicted secondary structure comprises of 1 α helix and 1 β strand.

6. 72 amino acid residue LND domain: The protein corresponding to GENE_ID ST0617 comprising of 373 amino acid residues consists of 72 amino acid residue region as four copies in tandem. Further PSI-BLAST searches corresponding to the region (19-90) as query identified four copies in *S. tokodaii* str. 7, *S. solfataricus* P2 and as three copies in tandem in *S. acidocaldarius* DSM 639 and *M. sedula* DSM 5348. The multiple sequence alignment identified LND as conserved sequence motif. The pair-wise sequence identities corresponding to the LND domain varied between 25-61%. The length of the proteins varied between 300 to 377 amino acid residues. The predicted secondary structure comprises of 4 α helices.

7. 100 amino acid residue GQP domain: The protein corresponding to GENE_ID ST1102 comprising of 895 amino acid residues consists of 100 amino acid residue region as two copies in tandem. Further PSI-BLAST searches corresponding to the region (21-120) as query identified two copies in tandem in *S. tokodaii* str. 7, *S. solfataricus* P2 and *S. acidocaldarius* DSM 639. The length of the proteins varied between 884 to 902 amino acid residues. The multiple sequence alignment identified GQP as conserved sequence motif. The pair-wise sequence identities corresponding to the GQP domain varied between 25-92%. The predicted secondary structure comprises of 1 α helix and 6 β strands. We observed that the GQP domain belongs to COG1449 (Cluster of Orthologues) and is predicted to function as sugar transporter permease protein.

8. 76 amino acid residue ExG domain: The protein corresponding to GENE_ID ST1658 comprising of 890 amino acid residues consists of 76 amino acid residue region as three copies in tandem. Further PSI-BLAST searches corresponding to the region (602-677) as query identified three copies in tandem in *S. tokodaii* str. 7, *T. volcanium* GSS1, four copies in *F. acidarmanus* Fer1 (GENE_ID FaciDRAFT_1608), two copies in tandem in *P. torridus* DSM 9790 and one copy in *F. acidarmanus* Fer1 (GENE_ID FaciDRAFT_0836), *M. sedula* DSM 5348 and *S. acidocaldarius* DSM 639. The multiple sequence alignment identified ExG as conserved sequence motif. The pair-wise sequence identities corresponding to the ExG domain varied between 12-46%. The length of the proteins varied between 514 to 972 amino acid residues. The predicted secondary structure comprises of 6 β strands.

9. 129 amino acid residue WTW domain: The protein corresponding to GENE_ID ST2253 comprising of 781 amino acid residues consists of 129 amino acid residue region as two copies. Further PSI-BLAST searches corresponding to the region (452-580) as query identified two copies in *S. tokodaii* str. 7, *S. acidocaldarius* DSM 639, *S. solfataricus* P2, *M. sedula* DSM 5348 and one copy in *P. torridus* DSM 9790. The multiple sequence alignment

identified WTW as conserved sequence motif. The pair-wise sequence identities corresponding to the WTW domain varied between 16-65%. The length of the proteins varied between 337 to 781 amino acid residues. The predicted secondary structure comprises of 4 β strands. This domain occurs along with the GTY repeat mentioned earlier.

10. 66 amino acid residue YPN domain: The protein corresponding to GENE_ID ST2364 comprising of 1301 amino acid residues consists of 66 amino acid residue region as three copies. Further PSI-BLAST searches corresponding to the region (579-644) as query identified three copies in *S. tokodaii* str. 7, *S. solfataricus* P2, two copies in *S. tokodaii* str. 7 (GENE_ID ST1692) and one copy in *S. acidocaldarius* DSM 639 and *M. sedula* DSM 5348. The multiple sequence alignment identified YPN as conserved sequence motif. The pair-wise sequence identities corresponding to the YPN domain varied between 7-59%. The length of the proteins varied between 1177 to 1308 amino acid residues. The predicted secondary structure comprises of 2 β strands.

11. 73 amino acid residue TYY domain: The protein corresponding to GENE_ID ST2475 comprising of 988 amino acid residues consists of 73 amino acid residue region as seven copies in tandem. Further PSI-BLAST searches corresponding to the region (412-484) as query identified variable copy numbers in tandem in 9 proteins. The multiple sequence alignment identified TYY as conserved sequence motif. The pair-wise sequence identities corresponding to the TYY domain varied between 11-74%. The length of the proteins varied between 548 to 1064 amino acid residues. The predicted secondary structure comprises of 5 β strands.

12. 68 amino acid residue GxL domain: The protein corresponding to GENE_ID ST2487 comprising of 284 amino acid residues consists of 68 amino acid residue region as three copies in tandem. Further PSI-BLAST searches

identified three copies in tandem in *S. tokodaii* str. 7, two copies in tandem in *S. solfataricus* P2, *P. torridus* DSM 9790, *M. sedula* DSM 5348 and *F. acidarmanus* Fer1. The multiple sequence alignment identified GxL as conserved sequence motif. The pair-wise sequence identities corresponding to the GxL domain varied between 22-70%. The length of the proteins varied between 196 to 284 amino acid residues. The predicted secondary structure comprises of 5 α helices.

The occurrence of these repeats and domains in this proteome is shown in Table 5c.

III. *Pyrobaculum aerophilum* str. IM2: *Pyrobaculum aerophilum* proteome comprises of 4 repeats and 5 domains.

1. 85 amino acid residue AAG domain: The 640 amino acid residues protein corresponding to the GENE_ID PAE0827 and described as hypothetical protein comprises of 85 amino acid residues region as three copies. Further PSI-BLAST searches using sequence corresponding to the region (89-173) as a query identified 2 proteins that are described as hypothetical proteins. This region occurs as three copies in proteins from *P. aerophilum* str. IM2 and *P. islandicum* DSM 4184. The length of the proteins varied from 638 to 640 amino acid residues. The multiple sequence alignment corresponding to this domain is associated with AAG sequence motif. The pair-wise identities between sequences corresponding to AAG domain varied between 12-62%. The secondary structure corresponding to AAG domain is predicted to comprise 2 β strands.

2. 72 amino acid residue GFGN domain: The protein corresponding to the GENE_ID PAE0829 comprising 2659 amino acid residues contains a novel 72 amino acid residues GFGN domain and a novel 43 amino acid residues KGG repeat. The 2659 amino acid residues protein corresponding to the GENE_ID PAE0829 and described as hypothetical protein comprises of a 72 amino acid

residues region as three copies. Further PSI-BLAST searches using sequence corresponding to the region (2376-2447) as a query identified 2 proteins that are described as hypothetical proteins. This region occurs as three copies in the proteins of *P. aerophilum* str. IM2 and *P. islandicum* DSM 4184. The length of the proteins varied from 2656 to 2659 amino acid residues. The multiple sequence alignment corresponding to this domain is associated with GFGN sequence motif. The pair-wise identities between sequences corresponding to GFGN domain varied between 33-76%. The secondary structure corresponding to GFGN domain is predicted to comprise 6 β strands.

3. 43 amino acid residue KGG repeat: The KGG repeat occurs as two copies in *P. aerophilum* str. IM2 corresponding to the region (2236-2278). The length of this region is less than 55 amino acid residues therefore, we refer this region as a repeat. The multiple sequence alignment corresponding to this repeat is associated with KGG sequence motif. The sequence homology shared between this KGG repeats is 34%. The secondary structure corresponding to KGG repeat is predicted to comprise 3 β strands.

4. 25 amino acid residue RWE repeat: The 411 amino acid residues protein corresponding to the GENE_ID PAE0906 and described as hypothetical protein comprises of a 25 amino acid residues region as ten copies. Further PSI-BLAST searches using sequence corresponding to the region (44-68) as a query did not identify any proteins in other organisms and therefore, is specific to *P. aerophilum* str. IM2 proteome. This region occurs as ten copies and in tandem, we therefore, describe this region as a repeat. The multiple sequence alignment corresponding to this repeat is associated with RWE sequence motif. The pair-wise identities between sequences corresponding to RWE repeats varied between 4-92%. The secondary structure corresponding to RWE repeat is predicted to comprise 1 α helix.

5. 25 amino acid residue RID repeat: The 191 amino acid residues protein corresponding to the GENE_ID PAE0920 and described as coiled-coil protein corresponding to the region (70-94) as a query comprises of a 25 amino acid residues region as three copies. This repeat is specific to *P. aerophilum* str. IM2 proteome. The multiple sequence alignment corresponding to this domain is associated with RID sequence motif. The pair-wise identities between sequences corresponding to RID repeat varied between 44-52%. The secondary structure corresponding to RID repeat is predicted to comprise 2 α helices.

6. 108 amino acid residue NDFA domain: The 221 amino acid residues protein corresponding to the GENE_ID PAE1277 and described as hypothetical protein comprises of a 108 amino acid residues region as two copies. Further PSI-BLAST searches using sequence corresponding to the region (33-140) as a query identified 12 proteins are described as hypothetical proteins. This region occurs as two copies in proteins from *P. aerophilum* str. IM2, *S. solfataricus* P2, *S. acidocaldarius* DSM 639, *S. tokodaii* str. 7, *T. tenax*, *P. carbinolicus* DSM 2380, *delta proteobacterium* MLMS-1, *delta proteobacterium* MLMS-1, *M. thermophila* PT, *M. sedula* DSM 5348, *S. fumaroxidans* MPOB and *S. aciditrophicus* SB. The length of these proteins varied between 214 to 251 amino acid residues. The multiple sequence alignment corresponding to this domain is associated with NDFA sequence motif. The pair-wise identities between sequences corresponding to NDFA domain varied between 22-99%. The secondary structure corresponding to NDFA domain is predicted to comprise 4 α helices and 3 β strands. The NDFA domain belongs to COG0438M and is predicted to function as trehalose-6-phosphate synthase.

7. 140 amino acid residue VxY domain: The 745 amino acid residues protein corresponding to the GENE_ID PAE1946 and described as hypothetical protein comprises of a 140 amino acid residues region as two copies. Further PSI-BLAST searches using sequence corresponding to the region (256-395) as a query identified 2 proteins that are described as hypothetical proteins. This

region occurs as two copies in proteins from *P. aerophilum* str. IM2 and *P. islandicum* DSM 4184. The length of the proteins varied between 745 to 1167 amino acid residues. The multiple sequence alignment corresponding to this domain is associated with VxY sequence motif. The pair-wise identities between sequences corresponding to VxY domain varied between 4-37%. The secondary structure corresponding to VxY domain is predicted to comprise 11 β strands.

8. 35 amino acid residue LLPN repeat: The 142 amino acid residues protein corresponding to the GENE_ID PAE3017 and described as hypothetical protein comprises of a 35 amino acid residue region as two copies. Further PSI-BLAST searches using sequence corresponding to the region (56-90) as a query identified 2 proteins that are described as hypothetical proteins. This region occurs as two copies in tandem in protein from *P. aerophilum* str. IM2 and three copies in protein from *P. islandicum* DSM 4184, we therefore, describe this region as a repeat. The length of the proteins varied between 142 to 264 amino acid residues. The multiple sequence alignment corresponding to this domain is associated with LLPN sequence motif. The pair-wise identities between sequences corresponding to LLPN repeat varied between 9-54%. The secondary structure corresponding to LLPN repeat is predicted to comprise 4 β strands.

9. 98 amino acid residue GxY domain: The 790 amino acid residues protein corresponding to the GENE_ID PAE3356 and described as hypothetical protein comprises of a 98 amino acid residues region as three copies. Further PSI-BLAST searches using sequence corresponding to the region (342-439) as a query identified 6 proteins that are described as hypothetical proteins. This region occurs as three copies in proteins from *P. aerophilum* str. IM2, *P. calidifontis* JCM 11548, *S. solfataricus* P2, *S. tokodaii* str.7, *S. acidocaldarius* DSM 639 and *M. sedula* DSM 5348. The length of the proteins varied from 646 to 974 amino acid residues. The multiple sequence alignment corresponding to this domain is associated with GxY sequence motif. The pair-

wise identities between sequences corresponding to GxY domain varied between 6-77%. The secondary structure corresponding to GxY domain is predicted to comprise 4 β strands. The occurrence of these repeats and domains is shown in Table 5d.

Euryarchaeota

(Archaeoglobus fulgidus* DSM 4304, *Halobacterium salinarium* NRC-1, *Methanobacterium thermoautotrophicum* str. Delta H, *Methanocaldococcus jannaschii* DSM 2661, *Methanopyrus kandleri* AV19, *Methanosarcina acetivorans* str. C2A, *Pyrococcus abyssi* GE5, *Thermoplasma acidophilum* DSM 1728, *Picrophilus torridus* DSM 9790*)

I. *Archaeoglobus fulgidus* DSM 4304*:* The proteome *Archaeoglobus fulgidus* comprises of 4 repeats and 7 domains.

1. 45 amino acid residue LIST repeat: The protein corresponding to the GENE_ID AF0002 comprising of 175 amino acid residues consists of 45 amino acid residue region as three copies in tandem. Further PSI-BLAST searches corresponding to the region (30-74) identified three copies in *A. fulgidus* DSM 4304 (GENE_ID AF0002) and two copies in *A. fulgidus* DSM 4304 (GENE_ID AF0443). The multiple sequence alignment identified LIST as conserved sequence motif. The pair-wise identities between sequences corresponding to LIST repeat varied between 23-53%. The length of the proteins varied from 114 to 175 amino acid residues. The secondary structure is predicted to comprise 2 β strands.

2. 41 amino acid residue GSY repeat: The protein corresponding to the GENE_ID AF0214 comprising of 676 amino acid residues consists of 41 amino acid residue repeat as well as a 67 amino acid residue SDL domain as two copies. The 41 amino acid residue GSY repeat corresponds to the region (380-420) as query identified two copies in *A. fulgidus* DSM 4304. The multiple sequence alignment identified GSY as conserved sequence motif. The sequence

homology shared between the GSY repeats is about 39%. The secondary structure is predicted to comprise 2 β strands.

3. 32 amino acid residue KEE repeat: The protein corresponding to the GENE_ID AF1557 comprising of 167 amino acid residues corresponding to the region (18-49) as query consists of 32 amino acid residue region as two copies. The multiple sequence alignment identified KEE as conserved sequence motif. The sequence homology shared between the KEE repeats is about 21%. The secondary structure is predicted to comprise 3 α helices.

4. 25 amino acid residue FQSP repeat: The protein corresponding to the GENE_ID AF1881 comprising of 247 amino acid residues consists of 25 amino acid residue region as eight copies in tandem. The multiple sequence alignment identified FQSP as conserved sequence motif. The pair-wise identities between sequences corresponding to FQSP repeats varied between 14-68%. The secondary structure is predicted to comprise 1 β strand.

5. 67 amino acid residue SDL domain: The protein corresponding to the GENE_ID AF0214 comprising of 676 amino acid residues consists of 67 amino acid residue region as two copies. The multiple sequence alignment identified SDL as conserved sequence motif. The sequence homology shared between the SDL domain is about 28%. The secondary structure is predicted to comprise 3 β strands.

6. 83 amino acid residue CCE domain: The protein corresponding to the GENE_ID AF0275 comprising of 914 amino acid residues consists of 83 amino acid residue region as four copies in tandem. Interestingly this protein has been described as a cell surface protein and we propose that these are cell surface protein specific domains. The multiple sequence alignment identified CCE as conserved sequence motif. The pair-wise identities between sequences corresponding to CCE domain varied between 23-68%. The secondary structure is predicted to comprise 5 β strands.

7. 93 amino acid residue EES domain: The protein corresponding to the GENE_ID AF1004 comprising of 522 amino acid residues consists of 93 amino acid residue region as two copies. The multiple sequence alignment identified EES as conserved sequence motif. The sequence homology shared between the EES domain is about 17%. The secondary structure is predicted to comprise 5 β strands.

8. 55 amino acid residue DGVL domain: The protein corresponding to the GENE_ID AF1820 comprising of 791 amino acid residues consists of 55 amino acid residue region as two copies. The multiple sequence alignment identified DGVL as conserved sequence motif. The sequence homology shared between the DGVL domain is about 37%. The secondary structure is predicted to comprise 2 α helices and 2 β strands.

9. 74 amino acid residue CPAGCE domain: The protein corresponding to the GENE_ID AF1948 comprising of 816 amino acid residues consists of 74 amino acid residue region as three copies. The multiple sequence alignment identified CPAGCE as conserved sequence motif. The pair-wise identities between sequences corresponding to CPAGCE domain varied between 38-47%. The secondary structure of the sequence is predicted to comprise mainly random coils.

10. 87 amino acid residue LAxY domain: The protein corresponding to the GENE_ID AF1994 comprising of 236 amino acid residues consists of 87 amino acid residue region as two copies. The multiple sequence alignment identified LAxY as conserved sequence motif. The sequence homology shared between the LAxY domain is about 22%. The secondary structure is predicted to comprise 1 α helix and 6 β strands.

11. 137 amino acid residue FxP domain: The protein corresponding to the GENE_ID AF2090-N comprising of 1948 amino acid residues consists of 137 amino acid residue region as three copies. Further PSI-BLAST searches with

the sequence corresponding to the region (30-166) as query identified three copies in *A. fulgidus* DSM 4304 (GENE_ID AF2090-N) and one copy in *M. jannaschii* DSM 2661 (GENE_ID MJ1396). The multiple sequence alignment identified FxP as conserved sequence motif. The pair-wise identities between sequences corresponding to FxP domain varied between 21-46%. The length of the proteins varied from 1948 to 2894 amino acid residues. The secondary structure is predicted to comprise 12 β strands.

The repeats and domains mentioned above are all specific to *A. fulgidus* DSM 4304 proteome except for FxP domain which also occurs in *M. jannaschii* DSM 2661 proteome. The occurrence of these repeats and domains in this proteome is shown in Table 5e.

II. *Halobacterium salinarium* NRC-1*:* The proteome *Halobacterium salinarium* NRC-1 comprises of 1 repeat and 8 domains.

The protein corresponding to the GENE_ID VNG0077H with length of 260 amino acid residues comprises of two domains, 1. 66 amino acid residue LxT domain, 2. 120 amino acid residue LEP domain and 3. 37 amino acid residue RxG repeat.

1. 37 amino acid residue RxG repeat: The region corresponding to (1-37) as query consists of 37 amino acid residue region as two copies in *H. salinarium* NRC-1 and one copy in *H. marismortui* ATCC 43049. The multiple sequence alignment identified RxG as conserved sequence motif. The pair-wise identities between sequences corresponding to RxG repeat varied between 32-56%. The length of the proteins varied between 260 to 547 amino acid residues. The secondary structure is predicted to comprise 3 β strands.

2. 66 amino acid residue LxT domain: The region corresponding to (38-103) as query consists of 66 amino acid residue region as one copy in *H. salinarium* NRC-1 and two copies in *H. marismortui* ATCC 43049. The multiple sequence alignment identified LxT as conserved sequence motif. The length of the

proteins varied between 260 to 547 amino acid residues. The pair-wise identities between sequences corresponding to LxT repeat varied between 27-45%. The secondary structure is predicted to comprise 5 β strands.

3. 120 amino acid residue LEP domain: The region corresponding to (141-260) as query consists of 120 amino acid residue region as one copy in *H. salinarium* NRC-1 and *H. marismortui* ATCC 43049. The multiple sequence alignment identified LEP as conserved sequence motif. The sequence homology shared between the LEP domain is about 60%. The length of the proteins varied between 260 to 547 amino acid residues. The secondary structure is predicted to comprise 1α helix and 9 β strands.

4. 62 amino acid residue GxW domain: The protein corresponding to the GENE_ID VNG7009 comprising of 772 amino acid residues consists of 62 amino acid residue region as nine copies in tandem. Further PSI-BLAST searches with the sequence corresponding to the region (6-67) as query identified nine copies in tandem in *H. salinarium* NRC-1 and three copies in tandem in *H. marismortui* ATCC 43049. The multiple sequence alignment identified GxW as conserved sequence motif. The pair-wise identities between sequences corresponding to GxW domain varied between 19-45%. The length of the proteins varied between 238 to 772 amino acid residues. The secondary structure is predicted to comprise 2 α helices.

5. 64 amino acid residue GxV domain: The protein corresponding to the GENE_ID VNG7113 comprising of 219 amino acid residues consists of 64 amino acid residue region as two copies. The multiple sequence alignment identified GxV as conserved sequence motif. The sequence homology shared between the GxV domain is about 28%. The secondary structure is predicted to comprise 5 β strands.

6. 55 amino acid residue SCT domain: The protein corresponding to the GENE_ID VNG0027H comprising of 215 amino acid residues consists of 55

amino acid residue region as two copies. The multiple sequence alignment identified SCT as conserved sequence motif. The sequence homology shared between the SCT domain is about 47%. The secondary structure is predicted to comprise 2 β strands.

7. 106 amino acid residue HExxE domain: The protein corresponding to the GENE_ID VNG0249G comprising of 810 amino acid residues consists of 106 amino acid residue region as five copies. Further PSI-BLAST searches corresponding to the region (63-167) as query identified five copies (2+3 tandem) in *H. salinarium* NRC-1 and in *H. marismortui* ATCC 43049. The multiple sequence alignment identified HExxE as conserved sequence motif. The pair-wise identities between sequences corresponding to HExxE domain varied between 11-52%. The length of the proteins varied between 810 to 823 amino acid residues. The secondary structure is predicted to comprise 6 α helices.

8. 58 amino acid residue PGE domain: The protein corresponding to the GENE_ID VNG1475C comprising of 551 amino acid residues consists of 58 amino acid residue region as three copies. The multiple sequence alignment identified PGE as conserved sequence motif. The sequence homology shared between the PGE domain is about 17%. The secondary structure is predicted to comprise 4 β strands.

9. 87 amino acid residue VxA domain: The protein corresponding to the GENE_ID VNG1953C comprising of 1363 amino acid residues consists of 87 amino acid residue region as three copies. The multiple sequence alignment identified VxA as conserved sequence motif. The pair-wise identities between sequences corresponding to VxA domain varied between 27-35%. The secondary structure is predicted to comprise 6 β strands.

The GxV, SCT, PGE and VxA domains are *H. salinarium* NRC-1 specific, RxG repeat, GxW, LxT, LEP, HExxE domains occurs in *H. salinarium* NRC-1 and

H. marismortui ATCC 43049. The occurrence of these repeats and domains in this proteome is shown in Table 5f.

III. *Methanobacterium thermoautotrophicum* str. Delta H: The *Methanobacterium thermoautotrophicum* proteome comprises of 2 repeats and 5 domains.

1. 48 amino acid residue RxP repeat: The protein corresponding to the GENE_ID MTH795 comprising of 405 amino acid residues consists of 48 amino acid residue region as two copies. The multiple sequence alignment identified RxP as conserved sequence motif. The sequence homology shared between the RxP repeat is about 29%. The secondary structure is predicted to comprise 1 β strand.

2. 45 amino acid residue YTxP repeat: The protein corresponding to the GENE_ID MTH910 comprising of 216 amino acid residues consists of 45 amino acid residue region as two copies. The multiple sequence alignment identified YTxP as conserved sequence motif. The sequence homology shared between the YTxP repeat is 35%. The secondary structure is predicted to comprise 5 β strands.

3. 66 amino acid residue VxV domain: The protein corresponding to the GENE_ID MTH179 comprising of 357 amino acid residues consists of 66 amino acid residue region as two copies in tandem. Further PSI-BLAST searches corresponding to the region (207-272) as query identified two copies in tandem in *M. thermautotrophicus* str. Delta H, *M. stadtmanae* DSM 3091 and *M. kandleri* AV19. The multiple sequence alignment identified VxV as conserved sequence motif. The pair-wise identities between sequences corresponding to VxV domain varied between 7-75%. The length of the proteins varied from 238 to 357 amino acid residues. The secondary structure is predicted to comprise 1 α helix and 2 β strands.

4. 115 amino acid residue CREC domain: The protein corresponding to the GENE_ID MTH309 comprising of 216 amino acid residues consists of 45 amino acid residue region as three copies (1+2 tandem). The multiple sequence alignment identified CREC as conserved sequence motif. The pair-wise identities between sequences corresponding to CREC domain varied between 22-25%. The secondary structure is predicted to comprise 3 α helices.

5. 187 amino acid residue CPG domain: The protein corresponding to the GENE_ID MTH674 comprising of 966 amino acid residues consists of 187 amino acid residue region as two copies. Further PSI-BLAST searches corresponding to the region (124-310) as query identified two copies in *M. thermautotrophicus* str. Delta H (GENE_ID MTH674), *M. barkeri* str. Fusaro (GENE_ID Mbar_A2934), *M. thermautotrophicus* str. Delta H (GENE_ID MTH1346), *M. mazei* Go1 (GENE_ID MM1875), *M. acetivorans* C2A (GENE_ID MA0715), *M. kandleri* AV19 (GENE_ID MK1177) and one copy in rest of the nine proteins. The multiple sequence alignment identified CPG as conserved sequence motif. The pair-wise identities between sequences corresponding to CPG domain varied between 8-94%. The length of the proteins varied from 361 to 966 amino acid residues. The secondary structure is predicted to comprise 4 α helices and 4 β strands.

6. 148 amino acid residue TPG domain: The TPG domain occurs along with CPG domain. The sequence corresponding to region (321-459) as query occurs as two copies in *M. thermautotrophicus* str. Delta H (GENE_ID MTH674), *M. barkeri* str. Fusaro (GENE_ID Mbar_A2934), *M. thermautotrophicus* str. Delta H (GENE_ID MTH1346), *M. mazei* Go1 (GENE_ID MM1875), *M. acetivorans* C2A (GENE_ID MA0715), *M. kandleri* AV19 (GENE_ID MK1177) and one copy in rest of the eight proteins. The multiple sequence alignment identified TPG as conserved sequence motif. The length of the proteins varied from 361

to 966 amino acid residues. The secondary structure corresponding to TPG domain is predicted to comprise 2 α helices and 5 β strands.

The RxP, YTxP repeats and CREC domain are *M. thermautotrophicus* str. Delta H specific, the VxV, CPG and TPG domains are seen in other archaeal genomes. The occurrence of these repeats and domains in this proteome is shown in Table 5g.

IV. *Methanocaldococcus jannaschii* DSM 2661: The *Methanocaldococcus jannaschii* DSM 2661 proteome comprises of 2 repeats and 5 domains.

1. 27 amino acid residue CG repeat: The protein corresponding to the GENE_ID MJ1230 comprising 76 amino acid residues consists of 27 amino acid residue region as two copies. Further PSI-BLAST searches corresponding to the region (7-33) as query identified two copies in *M. jannaschii* DSM 2661, *B. licheniformis* ATCC 14580, *O. iheyensis* HTE831 and *M. barkeri* str. Fusaro. The multiple sequence alignment identified CG as conserved sequence motif. The pair-wise identities between sequences corresponding to CG repeat varied between 13-70%. The length of the proteins varied from 69 to 76 amino acid residues. The secondary structure is predicted to comprise 2 β strands.

2. 31 amino acid residue CGA repeats: The protein corresponding to the GENE_ID MJ0409 comprising of 190 amino acid residues consists of 31 amino acid residue region as two copies. The multiple sequence alignment identified CGA as conserved sequence motif. The sequence homology shared between the CGA repeat is about 35%. The predicted secondary structure comprises of mainly loops.

3. 91 amino acid residue GYI domain: The protein corresponding to the GENE_ID MJ0164 comprising of 395 amino acid residues consists of 91 amino acid residue region as three copies. The multiple sequence alignment identified GYI as conserved sequence motif. The sequence homology shared between the

GYI domain is about 27%. The secondary structure is predicted to comprise 2 α helices and 2 β strands.

4. 58 amino acid residue IPDY domain: The protein corresponding to the GENE_ID MJ0409 comprising of 703 amino acid residues consists of 58 amino acid residue region as three copies. The multiple sequence alignment identified IPDY as conserved sequence motif. The sequence homology shared between the IPDY domain is about 27%. The secondary structure is predicted to comprise 3 β strands.

5. 90 amino acid residue IxE domain: The protein corresponding to the GENE_ID MJ0602 comprising of 261 amino acid residues consists of 90 amino acid residue region as three copies in tandem. The multiple sequence alignment identified IxE as conserved sequence motif. The pair-wise identities between sequences corresponding to IxE domain varied between 23-38%. The secondary structure is predicted to comprise 4 α helices.

6. 185 amino acid residue FYD domain: The protein corresponding to the GENE_ID MJ1394 comprising of 987 amino acid residues consists of 185 amino acid residue region as two copies. The multiple sequence alignment identified FYD as conserved sequence motif. The sequence homology shared between the FYD domain is about 56%. The secondary structure is predicted to comprise 1 α helix and 8 β strands.

7. 66 amino acid residue TLY domain: The protein corresponding to the GENE_ID MJ1584 comprising of 151 amino acid residues consists of 66 amino acid residue region as two copies. The multiple sequence alignment identified TLY as conserved sequence motif. The sequence homology shared between the TLY domain is about 51%. The secondary structure is predicted to comprise 4 α helices and 1 β strand.

The CGA repeat, GYI, IPDY, IxE, FYD and TLY domains are *M. jannaschii* DSM 2661 specific while the CG repeat is also seen in archaeal and bacterial

genomes. The occurrence of these repeats and domains in this genome is shown in Table 5h.

V. *Methanopyrus kandleri* AV19: The *Methanopyrus kandleri* AV19 proteome comprises of 2 domains.

1. 100 amino acid residue DWCA domain: The protein corresponding to the GENE_ID MK0947 comprising of 501 amino acid residues consists of 100 amino acid residue region as three copies in tandem. Further PSI-BLAST searches corresponding to the region (63-162) as query identified three copies in tandem in *M. kandleri* AV19 and one copy in *M. kandleri* AV19 (GENE_ID MK0948). The multiple sequence alignment identified DWCA as conserved sequence motif. The pair-wise identities between sequences corresponding to DWCA domain varied between 28-32%. The secondary structure is predicted to comprise 6 β strands.

2. 213 amino acid residue DYG domain: The protein corresponding to the GENE_ID MK1148 comprising of 1632 amino acid residues consists of 213 amino acid residue region as four copies in tandem. Further PSI-BLAST searches corresponding to the region (686-895) as query identified four copies in tandem in *M. kandleri* AV19 and one copy in *M. kandleri* AV19 (GENE_ID MK1149). The multiple sequence alignment identified DYG as conserved sequence motif. The pair-wise identities between sequences corresponding to DYG domain varied between 9-27%. The secondary structure is predicted to comprise 10 β strands.

The two domains DWCA and DYG are specific to *M. kandleri* AV19 proteome. The occurrence of the domains in this proteome is shown in Table 5i.

VI. *Methanosarcina acetivorans* str. C2A: The identification of novel repeats and domains in the cell surface proteins of *Methanosarcina acetivorans* str. C2A proteome was reported earlier (Adindla *et al*., 2004). However, the

analysis was carried out again using TRUST program in order to identify new repeats in the entire genome and we have identified 13 repeats and 8 domains. These are novel and have not been identified so far.

1. 24 amino acid residue KLK repeat: The protein corresponding to the GENE_ID MA2298 comprising of 111 amino acid residues consists of 24 amino acid residue region as three copies. The multiple sequence alignment identified KLK as conserved sequence motif. The sequence homology shared between the KLK repeat is about 45%. The secondary structure is predicted to comprise 1 α helix.

2. 28 amino acid residue SIV repeat: The protein corresponding to the GENE_ID MA2342 comprising of 187 amino acid residues consists of 28 amino acid residue region as two copies in tandem. Further PSI-BLAST searches corresponding to the region (21-48) as query identified two copies in tandem in *M. acetivorans* C2A, *M. mazei* Go1 and *M. barkeri* str. Fusaro. The multiple sequence alignment identified SIV as conserved sequence motif. The pair-wise identities between sequences corresponding to SIV repeat varied between 53-82%. The length of the proteins varied from 182 to 187 amino acid residues. The secondary structure is predicted to comprise 3 α helices.

3. 24 amino acid residue DDR repeat: The protein corresponding to the GENE_ID MA2713 comprising of 821 amino acid residues consists of 24 amino acid residue region as three copies in tandem. The multiple sequence alignment identified DDR as conserved sequence motif. The pair-wise identities between sequences corresponding to DDR repeat varied between 29-62%. The secondary structure is predicted to comprise 1 α helix.

4. 17 amino acid residue TQN repeat: The protein corresponding to the GENE_ID MA2913 comprising of 108 amino acid residues consists of 17 amino acid residue region as five copies in tandem. The multiple sequence alignment identified TQN as conserved sequence motif. The pair-wise

identities between sequences corresponding to TQN repeat varied between 82-100%. The secondary structure is predicted to comprise 1 α helix.

5. 36 amino acid residue PxL repeat: The protein corresponding to the GENE_ID MA3387 comprising of 417 amino acid residues consists of 36 amino acid residue region as four copies in tandem. Further PSI-BLAST searches corresponding to the region (181-216) as query identified four copies in tandem in *M. acetivorans* C2A, *M. mazei* Go1 (GENE_ID MM2676), *M. barkeri* str. Fusaro and six copies in *M. mazei* Go1 (GENE_ID MM3244). The multiple sequence alignment identified PxL as conserved sequence motif. The pair-wise identities between sequences corresponding to PxL repeat varied between 25-72%. The length of the proteins varied from 417 to 869 amino acid residues. The secondary structure is predicted to comprise 2 α helices.

6. 43 amino acid residue ELI repeat: The protein corresponding to the GENE_ID MA3812 comprising of 242 amino acid residues consists of 43 amino acid residue region as three copies. The multiple sequence alignment identified ELI as conserved sequence motif. The pair-wise identities between sequences corresponding to ELI repeat varied between 2-34%. The secondary structure is predicted to comprise 1 α helix and 1 β strand.

7. 49 amino acid residue LVC repeat: The protein corresponding to the GENE_ID MA0396 comprising of 114 amino acid residues consists of 49 amino acid residue region as two copies in tandem. Further PSI-BLAST searches corresponding to the region (3-52) as query identified two copies in tandem in *M. acetivorans* C2A, *M. barkeri* str. Fusaro (GENE_ID Mbar_A1221), *M. barkeri* str. Fusaro (GENE_ID Mbar_A1214) and one copy in *M. barkeri* str. Fusaro (GENE_ID Mbar_A1215). The multiple sequence alignment identified LVC as conserved sequence motif. The pair-wise identities between sequences corresponding to LVC repeat varied between 44-

92%. The length of the proteins varied from 59 to 114 amino acid residues. The secondary structure is predicted to comprise 1α helix and 3 β strands.

8. 46 amino acid residue NLE repeat: The protein corresponding to the GENE_ID MA1577 comprising of 401 amino acid residues consists of 46 amino acid residue region as seven copies in tandem. Further PSI-BLAST searches corresponding to the region (110-155) as query identified seven copies in tandem in *M. acetivorans* C2A (GENE_ID MA1577), four copies in tandem in *M. acetivorans* C2A (GENE_ID MA1580), fourteen copies in tandem in *M. barkeri* str. Fusaro (GENE_ID Mbar_A2800) and (2 tandem + 1) in *M. barkeri* str. Fusaro (GENE_ID Mbar_A2801). The multiple sequence alignment identified NLE as conserved sequence motif. The pair-wise identities between sequences corresponding to NLE repeat varied between 17-89%. The length of the proteins varied from 341 to 720 amino acid residues. The secondary structure is predicted to comprise 4 α helices.

9. 42 amino acid residue GE repeat: The protein corresponding to the GENE_ID MA1641 comprising of 668 amino acid residues consists of 42 amino acid residue region as three copies. The multiple sequence alignment identified GE as conserved sequence motif. The pair-wise identities between sequences corresponding to GE repeat varied between 7-71%. The secondary structure is predicted to comprise 2 β strands.

10. 36 amino acid residue NLG repeat: The protein corresponding to the GENE_ID MA1785 comprising of 156 amino acid residues consists of 36 amino acid residue region as three copies. The multiple sequence alignment identified NLG as conserved sequence motif. The sequence homology shared between the NLG repeat is about 52%. The secondary structure is predicted to comprise 3 α helices.

11. 23 amino acid residue FNP repeat: The protein corresponding to the GENE_ID MA1927 comprising of 141 amino acid residues corresponding to

the region (5-27) as query consists of 23 amino acid residue region as three copies (1+ 2 tandem). The multiple sequence alignment identified FNP as conserved sequence motif. The pair-wise identities between sequences corresponding to FNP repeat varied between 45-52%. The secondary structure is predicted to comprise 1 α helix.

12. 26 amino acid residue WVP repeat: The protein corresponding to the GENE_ID MA1984 comprising of 127 amino acid residues corresponding to the region (64-89) as query consists of 26 amino acid residue region as two copies. The multiple sequence alignment identified WVP as conserved sequence motif. The sequence homology shared between the WVP repeat is about 46%. The secondary structure is predicted to comprise 2 β strands.

13. 58 amino acid residue PLM domain: The protein corresponding to the GENE_ID MA2106 comprising of 955 amino acid residues consists of 58 amino acid residue region as two copies. Further PSI-BLAST searches corresponding to the region (606-663) as query identified two copies in *M. acetivorans* C2A, *M. mazei* Go1 and *M. marisnigri* JR1. The length of the proteins varied from 945 to 955 amino acid residues. The multiple sequence alignment identified PLM as conserved sequence motif. The pair-wise identities between sequences corresponding to PLM domain varied between 24-98%. The secondary structure is predicted to comprise 3 α helices.

14. 135 amino acid residue LSW domain: The protein corresponding to the GENE_ID MA2307 comprising of 858 amino acid residues consists of 135 amino acid residue region as two copies. Further PSI-BLAST searches corresponding to the region (119-253) as query identified two copies in *M. acetivorans* C2A and *J. sp*. HTCC2649 (GENE_ID JNB_14143). The multiple sequence alignment identified LSW as conserved sequence motif. The pair-wise identities between sequences corresponding to LSW domain varied

between 28-49%. The length of the proteins varied between 734 to 858 amino acid residues. The secondary structure is predicted to comprise 4 β strands.

15. 85 amino acid residue STS domain: The protein corresponding to the GENE_ID MA2325 comprising of 719 amino acid residues corresponding to the region (15-99) as query consists of 85 amino acid residue region as two copies in tandem. The multiple sequence alignment identified STS as conserved sequence motif. The sequence homology shared between the STS domain is about 61%. The secondary structure is predicted to comprise 3 α helices and 2 β strands.

16. 232 amino acid residue GLW domain: The protein corresponding to the GENE_ID MA2713 comprising of 821 amino acid residues consists of 232 amino acid residue region as two copies. Further PSI-BLAST searches corresponding to the region (19-250) as query identified two copies in *M. acetivorans* C2A and *M. barkeri* str. Fusaro. The multiple sequence alignment identified GLW as conserved sequence motif. The pair-wise identities between sequences corresponding to GLW domain varied between 28-64%. The length of the proteins varied between 470 to 821 amino acid residues. The secondary structure is predicted to comprise 2 α helices and 5 β strands. This domain occurs along with 24 amino acid residue DDR repeats within the same proteins.

17. 36 amino acid residue KPE repeat: The protein corresponding to the GENE_ID MA4346 comprising of 183 amino acid residues consists of 36 amino acid residue region as two copies. Further PSI-BLAST searches corresponding to the region (6-41) as query identified two copies in *M. acetivorans* C2A, *M. mazei* Go1 and three copies in *M. barkeri* str. Fusaro. The multiple sequence alignment identified KPE as conserved sequence motif. The pair-wise identities between sequences corresponding to KPE repeat varied between 22-88%. The length of the proteins varied from 168 to 240 amino acid

residues. The predicted secondary structure of sequence comprises of mainly loops.

18. 103 amino acid residue GGY domain: The protein corresponding to the GENE_ID MA0163 comprising of 307 amino acid residues consists of 103 amino acid residue region as two copies. Further PSI-BLAST searches corresponding to the region (14-116) as query identified two copies in *M. acetivorans* C2A, *M. barkeri* str. Fusaro, *M. burtonii* DSM 6242, *H. marismortui* ATCC 43049, *H. salinarium* NRC-1 and *M. marisnigri* JR1. The multiple sequence alignment identified GGY as conserved sequence motif. The pair-wise identities between sequences corresponding to GGY domain varied between 13-72%. The length of the proteins varied from 307 to 334 amino acid residues. The secondary structure is predicted to comprise 5 β strands.

19. 133 amino acid residue AIK domain: The protein corresponding to the GENE_ID MA1936 comprising of 1078 amino acid residues consists of 133 amino acid residue region as two copies. Further PSI-BLAST searches corresponding to the region (371-503) as query identified two copies in *M. acetivorans* C2A and *M. barkeri* str. Fusaro. The multiple sequence alignment identified AIK as conserved sequence motif. The pair-wise identities between sequences corresponding to AIK domain varied between 18-56%. The length of the proteins varied between 1078 to 1123 amino acid residues. The secondary structure is predicted to comprise 7 α helices.

20. 92 amino acid residue GxD domain: The protein corresponding to the GENE_ID MA1841 comprising of 628 amino acid residues corresponding to the region (116-207) as query consists of 92 amino acid residue region as three copies in tandem. The multiple sequence alignment identified GxD as conserved sequence motif. The pair-wise identities between sequences corresponding to GxD domain varied between 6-32%. The secondary structure is predicted to comprise 4 β strands.

21. 57 amino acid residue YP domain: The protein corresponding to the GENE_ID MA2331 comprising of 95 amino acid residues consists of 57 amino acid residue region as one copy. Further PSI-BLAST searches corresponding to the region (13-69) as query identified one, two and three copies in 68 proteins from various archaeal and bacterial genomes. The domain occurs as three copies in *M. marisnigri* JR1 (GENE_ID Memar_0389), as two copies in *M. marisnigri* JR1 (GENE_ID Memar_0390), tandem in *M. barkeri* str. Fusaro (GENE_ID Mbar_A0068), *M. acetivorans* C2A (GENE_ID MA2331), *M. barkeri* str. Fusaro (GENE_ID Mbar_A1083) and *N. punctiforme* PCC 73102 (GENE_ID Npun02004635). The multiple sequence alignment identified YP as conserved sequence motif. The pair-wise identities between sequences corresponding to YP domain varied between 10-100%. The length of the proteins varied from 61 to 526 amino acid residues. The secondary structure is predicted to comprise 4 α helices.

The KLK, DDR, TQN, ELI, GE, NLG, FNP, WVP repeats and STS, GxD domains are specific to *M. acetivorans* C2A proteome, the PxL, LVC, NLE, SIV, KPE repeats and PLM, GLW, GGY and AIK domains are present in other archaeal genomes while LSW and YP domains are also seen in archaeal as well as bacterial genomes. The occurrence of these repeats and domains in this proteome is shown in Table 5j.

VII. *Pyrococcus abyssi* GE5: The proteome of *Pyrococcus abyssi* GE5 comprises of 4 domains.

1. 146 amino acid residue GYS domain: The protein corresponding to the GENE_ID PAB1860 comprising of 266 amino acid residues consists of 146 amino acid residue region as one copy. Further PSI-BLAST searches corresponding to the region (6-151) as query identified two copies in tandem in *P. horikoshii* OT3, *T. kodakarensis* KOD1 and *P. furiosus* DSM 3638. The multiple sequence alignment identified GYS as conserved sequence motif. The

pair-wise identities between sequences corresponding to GYS domain varied between 7-73%. The length of the proteins varied from 266 to 293 amino acid residues. The secondary structure is predicted to comprise 4 α helices and 3 β strands.

2. 59 amino acid residue GxF domain: The protein corresponding to the GENE_ID PAB1294 comprising of 595 amino acid residues consists of 59 amino acid residue region as one copy. Further PSI-BLAST searches corresponding to the region (312-370) as query identified one copy in *P. abyssi* GE5, *P. furiosus* DSM 3638 and two copies in *P. horikoshii* OT3. The multiple sequence alignment identified GxF as conserved sequence motif. The pair-wise identities between sequences corresponding to GxF domain varied between 18-46%. The length of the proteins varied from 562 to 633 amino acid residues. The secondary structure is predicted to comprise 5 β strands.

3. 56 amino acid residue VTI domain: The protein corresponding to the GENE_ID PAB1252 comprising of 1204 amino acid residues consists of 56 amino acid residue region as one copy. Further PSI-BLAST searches corresponding to the region (826-882) as query identified one copy in *P. abyssi* GE5, *P. horikoshii* OT3 and *T. kodakarensis* KOD1. The multiple sequence alignment identified VTI as conserved sequence motif. The pair-wise identities between sequences corresponding to VTI domain varied between 33-53%. The length of the proteins varied from 1103 to 1204 amino acid residues. The secondary structure is predicted to comprise 6 β strands.

4. 100 amino acid residue NxG domain: The protein corresponding to the GENE_ID PAB1102 comprising of 899 amino acid residues consists of 100 amino acid residue region as three copies. Further PSI-BLAST searches corresponding to the region (291-391) as query identified three copies in *P. abyssi* GE5, *P. horikoshii* OT3 and *P. furiosus* DSM 3638, two copies in Uncultured archaeon GZfos26B2 (GENE_ID GZ26B2_6), Uncultured archaeon

GZfos28B8 (GENE_ID GZ28B8_10) and one copy in *M. hungatei* JF-1 and *M. jannaschii* DSM 2661. The multiple sequence alignment identified NxG as conserved sequence motif. The pair-wise identities between sequences corresponding to NxG domain varied between 8-83%. The length of the proteins varied from 384 to 1155 amino acid residues. The secondary structure is predicted to comprise 5 β strands.

The domains mentioned above in *P. abyssi* GE5 are present in other archaeal genomes. The occurrence of these domains in this proteome is shown in Table 5k.

VIII. *Thermoplasma acidophilum* DSM 1728*:* The proteome of *Themoplsma acidophilum* DSM 1728 comprises of 4 domains. They are as follows:

1. 77 amino acid residue GLP domain: The protein corresponding to the GENE_ID Ta0167 comprising of 998 amino acid residues consists of 77 amino acid residue region as three copies in tandem. Further PSI-BLAST searches corresponding to the region (739-815) as a query identified three copies in tandem in *T. acidophilum* DSM 1728 and *T. volcanium* GSS1. The multiple sequence alignment identified GLP as conserved sequence motif. The pair-wise identities between sequences corresponding to GLP domain varied between 18-40%. The length of the proteins is 998 amino acid residues respectively. The secondary structure is predicted to comprise 7 β strands.

2. 69 amino acid residue GxY domain: The protein corresponding to the GENE_ID Ta0543 comprising of 1124 amino acid residues corresponding to the region (682-750) as query consists of 69 amino acid residue region as five copies in tandem. The multiple sequence alignment identified GxY as conserved sequence motif. The pair-wise identities between sequences corresponding to GxY domain varied between 8-31%. The secondary structure is predicted to comprise 4 β strands.

3. 91 amino acid residue IK domain: The protein corresponding to the GENE_ID Ta0587 comprising of 1690 amino acid residue consists of 91 amino acid residue region as one copy. Further PSI-BLAST searches corresponding to the region (1149-1239) as query identified one copy in *T. acidophilum* DSM 1728, *F. acidarmanus* fer1, *T. volcanium* GSS1 and *P. torridus* DSM 9790. The multiple sequence alignment identified IK as conserved sequence motif. The pair-wise identities between sequences corresponding to IK domain varied between 24-53%. The length of the proteins varied from 1667 to 1713 amino acid residues. The secondary structure is predicted to comprise 4 α helices and 2 β strands.

4. 167 amino acid residue TxN domain: The protein corresponding to the GENE_ID Ta1136 comprising of 2081 amino acid residues consists of 167 amino acid residue region as two copies in tandem. Further PSI-BLAST searches corresponding to the region (1904-2070) as query identified two copies in tandem in *T. acidophilum* DSM 1728, *P. torridus* DSM 9790, *T. volcanium* GSS1 and *F. acidarmanus* Fer1. The multiple sequence alignment identified TxN as conserved sequence motif. The pair-wise identities between sequences corresponding to TxN domain varied between 6-58%. The length of the proteins varied from 1637 to 2081 amino acid residues. The secondary structure is predicted to comprise 6 β strands.

The GxY domain is specific to *T. acidophilum* DSM 1728 proteome, the GLP domain is present in *T. acidophilum* DSM 1728 and *T. volcanium* GSS1 proteomes, while the IK domain and TxN domain are present in other archaeal genomes.

The occurrence of these domains in this proteome is shown in Table 5l.

IX.*Picrophilus torridus* DSM 9790:

The proteome of *Picrophilus torridus* DSM 9790 comprises of 6 repeats. They are as follows:

1. 30 amino acid residue YxxxG repeat: The protein corresponding to the GENE_ID PTO0099 comprising of 527 amino acid residues corresponding to the region (319-350) as query consists of 30 amino acid residue region as two copies. The multiple sequence alignment identified YxxxG as conserved sequence motif. The sequence homology shared between the YxxxG repeat is about 13%. The secondary structure is predicted to comprise 3 β strands.

2. 51 amino acid residue IYQ repeat: The protein corresponding to the GENE_ID PTO0352 comprising of 152 amino acid residues corresponding to the region (11-61) as query consists of 51 amino acid residue region as two copies. The multiple sequence alignment identified IYQ as conserved sequence motif. The sequence homology shared between the IYQ repeat is about 35%. The secondary structure is predicted to comprise 1 α helix and 3 β strands.

3. 51 amino acid residue YKL repeat: The protein corresponding to the GENE_ID PTO0786 with length of 1637 amino acid residues corresponding to the region (788-838) as query consists of 51 amino acid residue region as two copies. The multiple sequence alignment identified YKL as conserved sequence motif. The sequence homology shared between the YKL repeat is about 25%. The secondary structure is predicted to comprise 5 β strands.

4. 44 amino acid residue NNT repeat: The protein corresponding to the GENE_ID PTO0798 comprising of 546 amino acid residues consists of 44 amino acid residue region as two copies. Further PSI-BLAST searches corresponding to the region (362-403) as query identified two copies in *P. torridus* DSM 9790 and one copy in *T. volcanium* GSS1 (GENE_ID TVG0507890), *T. volcanium* GSS1 (GENE_ID TVN0518), *T. acidophilum* DSM 1728 and *F. acidarmanus* Fer1. The multiple sequence alignment identified NNT as conserved sequence motif. The pair-wise identities between sequences corresponding to NNT repeat varied between 7-100%. The length of

the proteins varied from 172 to 546 amino acid residues. The secondary structure is predicted to comprise 3 β strands.

5. 40 amino acid residue AW repeat: The protein corresponding to the GENE_ID PTO0842 comprising of 242 amino acid residues consists of 40 amino acid residue region as six copies. Further PSI-BLAST searches corresponding to the query (4-43) identified six copies in tandem in *P. torridus* DSM 9790 and *F. acidarmanus* Fer1. The multiple sequence alignment identified AW as conserved sequence motif. The pair-wise identities between sequences corresponding to AW repeat varied between 11-82%. The length of the proteins varied from 242 to 243 amino acid residues. The secondary structure is predicted to comprise 2 α helices.

6. 42 amino acid residue YN repeat: The protein corresponding to the GENE_ID PTO1487 comprising of 493 amino acid residues corresponding to the region (314-355) as query identified 42 amino acid residue region as three copies. The multiple sequence alignment identified YN as conserved sequence motif. The pair-wise identities between sequences corresponding to YN repeat varied between 11-16%. The secondary structure is predicted to comprise 4 β strands.

The YxxxG, IYQ, YKL and YN repeats are *P. torridus* DSM 9790 specific while the NNT and AW repeats are present in other archaeal genomes.

The occurrence of these repeats and domain in this proteome is shown in Table 5m.

Nanoarchaeota

(*Nanoarchaeum equitans* Kin4-M)

I. *Nanoarchaeum equitans* Kin4-M: The proteome of *Nanoarchaeum equitans* Kin4-M comprises of 1 repeat and 1 domain.

1. 33 amino acid residue YGGK repeat: The protein corresponding to the GENE_ID NEQ221 comprising of 98 amino acid residues corresponding to the region (12-44) as query consists of 33 amino acid residue as two copies. The multiple sequence alignment identified YGGK as conserved sequence motif. The sequence homology shared between the YGGK repeat is about 7%. The secondary structure is predicted to comprise of 1 β strand.

2. 55 amino acid residue DxLN domain: The protein corresponding to the GENE_ID NEQ032 comprising of 383 amino acid residues corresponding to the region (58-112) as query consists of 55 amino acid region as two copies in tandem. The multiple sequence alignment identified DxLN as conserved sequence motif. The sequence homology shared between the DxLN domain is about 36%. The secondary structure is predicted to comprise 2 α helices and 1 β strand. The repeat and domain mentioned above are specific to *N. equitans* Kin4-M proteome. The occurrence of these repeat and domain in this proteome is shown in Table 5n.

This exhaustive study to identify novel repeats and domains in archaeal genomes, throws light on the diversity of these organisms that can be broadly categorized as extermophiles. *A. pernix* K1 is a thermophile, with optimal growth at 90-95^{o}C, pH 5-9 and a salinity of 3.5%. We identified only 1 novel domain in an orphan protein. *S. tokodaii* str. 7 is a hyperthermophilic, acidophilic, sulfur-metabolizing archeon. We identified 7 domains and 5 repeats in proteins of which, 2 are orphans, 9 are archaeal specific and 1 protein is present in both archaeal and bacterial proteomes. *P. aerophilum* str. IM2 is a nitrate-reducing hyperthermophilic archeon. We identified 4 repeats and 5

domains in proteins of which, 3 are orphans, 2 are archaeal specific and 4 proteins are within the same genus.

A. fulgidus DSM 4304 is a sulfur-metabolizing organism and can grow at extremely high temperatures. We identified 4 novel repeats and 7 domains in proteins, of which, 10 are orphans and 1 protein is present within this sub-domain of archaea. *H. salinarium* str. NRC-1 is a halophilic archeon and is adapted to grow under extremely high saline conditions. We identified 1 repeat and 8 domains in proteins, of which, 4 are orphans and 5 proteins are within the same genus. *M. thermautotrophicus* str. Delta H is a thermophilic, methane producing archaea. We identified 2 repeats and 4 domains in proteins, of which, 3 are orphans and 3 proteins are present in other archaeal proteomes. *M. jannaschii* DSM 2661 is a methane-producing archaea. We identified 2 repeats and 5 domans in proteins, of which, 6 are orphans and 1 protein is present in both archaeal and bacterial proteomes. *M. kandleri* str. AV19 is a hyperthermophilic methanogen. We identified 2 domains in proteins that are orphans. *M. acetivorans* str. C2A is a non-motile, methane-producing archaea. We identified 13 repeats and 8 domains in proteins, of which, 10 are orphans, 9 are present in archaeal proteomes, 1 is present in bacteria and 1 protein is present in both archaeal and bacterial proteomes. *P. abyssi* str. GE5 is a hyperthermophilic archeon with optimal growth at 103°C at 200 atmospheres of pressure. We identified 4 domains in proteins that are present in other archaeal proteomes. *T. acidophilum* str. DSM 1728 is thermophilic and acidophilic. We identified 4 domains in proteins, of which 1 is orphan, 2 are present in other archaeal genomes and 1 is present in proteins within the same genus. *P. torridus* str. DSM 9790 is a thermoacidophile. This organism expresses a surface layer (S-layer) that consists of a semicrystalline array of proteins outside the cell membrane. We identified 6 repeats in proteins, of which 4 are orphans, 1 is present in other archaeal genomes and 1 protein is present within the same sub-domain of archaea.

N. equitans Kin4-M is a tiny microbe and requires a host for survival and has lost vital genes for several metabolic pathways. These features of this newly discovered genome explain the absence of known repeats or domains. We identified only 1 novel repeat and 1 domain in proteins that are orphans.

From the repeats and domains present in these representative archaeal genomes (based on the data previously known and the findings in this work, see table 5a), we make the follwing inferences. WD-40, NHL, LVIVD repeats fold into a 3-D beta-propeller architecture and are known to be present in the cell surface proteins of various bacterial and archaeal organisms. The presence of WD-40 repeats in *A. pernix* K1, *S. tokodaii* str. 7, *P. aerophilum* str. IM2, *H. salinarium* NRC-1, *M. acetivorans* C2A, *P. abyssi* GE5, *T. acidophilum* DSM 1728 and *P. torridus* DSM 9790, NHL repeats in *S. tokodaii* str. 7, *M. acetivorans* C2A, *T. acidophilum* DSM 1728 and *P. torridus* DSM 9790, and LVIVD repeats in *M. thermoautotrophicum* str. Delta H, *M. jannaschii* DSM 2661 and *M. acetivorans* C2A. indicate that these organisms consists of semicrystalline array of proteins outside their cell membrane.

The organism, *N. equitans* Kin4-M has least number of proteins and only one repeat and one domain in orphan proteins. Also exchange of genes within archaea are least observed in this proteome and we therefore propose that *N. equitans* Kin4-M is a minimalist archaea. The exchange of genes between archaeal and bacterial genomes is maximal in *M. acetivorans* C2A. The number of orphan proteins comprising repeats and domains is also high, indicating the extensive evolution of these genomes. This is required for their adaptation to extreme living conditions such as high temperature, pressure and pH.

Table 5a. Table showing the archaeal organisms, total number of protein sequences in the respective genomes, the number of repeat sequences identified by TRUST, types of known repeats, their presence and absence in the respective genomes, number of novel repeats and number of novel domains.

Organism	*A.. pern*	*S. toko*	*P. aero*	*A. fulg*	*H. sali*	*M. ther*	*M. jann*	*M. kand*	*M. acet*	*P. abys*	*T. acid*	*P. torri*	*N. equi*
Total no. of protein seq uences in proteome	1841	2825	2605	2420	2622	1873	1786	1687	4540	1896	1482	1535	536
No. of repe-at sequences identified by TRUST	310	462	418	443	515	380	398	327	1059	338	238	268	83
Known Repeats													
EZ-HEAT	×	×	√	×	√	√	×	×	√	√	×	×	×
BNR	√	×	√	×	√	√	√	×	√	×	×	×	×
TPR	×	√	×	√	√	√	√	×	√	√	√	√	×
PbH1	×	√	√	√	√	√	√	×	√	×	√	√	×
PQQ	×	√	×	√	√	√	√	×	√	√	×	√	×
FG-GAP	×	×	×	√	×	×	×	√	√	×	×	×	×
SBBP	×	×	×	×	×	√	√	×	√	×	×	×	×
Hexapeptide	√	√	√	×	√	√	√	√	√	√	√	√	×
PD40 /WD40	√	√	√	×	√	×	×	×	√	√	√	√	×
NHL	×	√	×	×	×	×	×	×	√	×	√	√	×
Pentapeptide	×	×	×	×	×	×	×	×	√	×	×	×	×
LVIVD	×	×	×	×	×	√	√	×	√	×	×	×	×
LGFP	×	×	×	×	×	×	×	×	√	×	×	×	×
ARM	×	×	√	×	√	√	×	×	√	×	×	×	×
LRR	×	×	×	×	×	×	×	×	√	×	×	×	×
DNA-gyrase	×	×	×	√	√	×	×	×	√	×	√	√	×
Ankyrin	×	×	√	×	×	×	×	×	×	×	√	×	×
Kelch	×	√	×	√	×	×	×	×	×	×	×	×	×
No. of novel Repeats	0	5	4	4	1	2	2	0	13	0	0	6	1
No. of novel Domains	1	7	5	7	8	4	5	2	8	4	4	0	1

"√ " indicates the presence of known repeats according to SMART nomenclature, "×" indicates the absence of the corresponding repeat. "*A. pern"* denotes *Aeropyrum pernix* K1, "*S. toko"* denotes *Sulfolobus tokodaii* str. 7, "*P. aero*" denotes *Pyrobaculum aerophilum* str. IM2, "*A. fulg*" denotes *Archaeoglobus fulgidus* DSM 4304, "*H. sali*" denotes *Halobacterium salinarium* NRC-1, "*M. ther*" denotes *Methanobacterium thermoautotrophicum* str. Delta H, "*M. jann*" denotes *Methanocaldococcus jannaschii* DSM 2661, "*M. kand*" denotes *Methanopyrus kandleri* AV19, "*M. acet*" denotes *Methanosarcina acetivorans* C2A, "*P. abys*" denotes *Pyrococcus abyssi* GE5, "*T. acid*" denotes *Thermoplasma acidophilum* DSM 1728, "*P. torr*" denotes *Picrophilus torridus* DSM 9790 and "*N. equi*" denotes *Nanoarchaeum equitans* Kin4-M.

Table 5b. Total number of novel repeats/domains in *Aeropyrum pernix* K1 proteome.

S. No	Repeat / Domain name	Length of Repeat / Domain (amino acids)	Predicted Secondary structure	Number of Proteins identified from PSI-BLAST	Taxonomy of repeat occurrence
1	PxG (D)	90	8 β strands	1	*Aeropyrum pernix* specific

Table 5c. Total number of novel repeats/domains in *Sulfolobus tokodaii* str. 7 proteome.

S. No	Repeat / Domain name	Length of Repeat / Domain (amino acids)	Predicted Secondary structure	Number of Proteins identified from PSI-BLAST	Taxonomy of repeat occurrence
1	LND (D)	72	4 α helices	4	*Sulfolobus tokodaii* and other Archaeal genomes
2	EYL (R)	46	βαββ	1	*Sulfolobus tokodaii* specific
3	GQP (D)	100	βββαβββ	3	*Sulfolobus tokodaii* and other Archaeal genomes
4	LVVV (R)	44	βαβ	4	*Sulfolobus tokodaii* and other Archaeal & Bacterial genomes
5	ExG (D)	76	6 β strands	7	*Sulfolobus tokodaii* and other Archaeal genomes
6	LIN (R)	30	4 α helices	4	*Sulfolobus tokodaii* and other Archaeal genomes
7	KxK (R)	43	ααββ	1	*Sulfolobus tokodaii* specific
8	WTW (D)	129	4 β strands	5	*Sulfolobus tokodaii* and other Archaeal genomes
9	GTY (R)	48	αβ	5	*Sulfolobus tokodaii* and other Archaeal genomes
10	YPN (D)	66	2 β strands	5	*Sulfolobus tokodaii* and other Archaeal genomes
11	TYY(D)	73	5 β strands	9	*Sulfolobus tokodaii* and other Archaeal genomes
12	GxL (D)	68	5 α helices	5	*Sulfolobus tokodaii* and other Archaeal genomes

Table 5d. Total number of novel repeats/domains in *Pyrobaculum aerophilum* str. IM2 proteome.

S. No	Repeat / Domain name	Length of Repeat / Domain (amino acids)	Predicted Secondary structure	Number of Proteins identified from PSI-BLAST	Taxonomy of repeat occurrence
1	AAG (D)	85	2 β strands	2	*Pyrobaculum aerophilum* and *Pyrobaculum islandicum* DSM 4184
2	GFGN (D)	72	6 β strands	2	*Pyrobaculum aerophilum* and *Pyrobaculum islandicum* DSM 4184
3	KGG (R)	43	3 β strands	1	*Pyrobaculum aerophilum* specific
4	RWE (R)	25	1 α helix	1	*Pyrobaculum aerophilum* specific
5	RID (R)	25	2 α helices	1	*Pyrobaculum_aerophilum* specific
6	NDFA (D)	108	αβαααββ	12	*Pyrobaculum aerophilum* and other Archaeal genomes
7	VxY (D)	140	11β strands	2	*Pyrobaculum aerophilum* and *Pyrobaculum islandicum* DSM 4184
8	LLPN (R)	35	4 β strands	2	*Pyrobaculum aerophilum* and *Pyrobaculum islandicum* DSM 4184
9	GxY (D)	98	4 β strands	6	*Pyrobaculum aerophilum* and other Archaeal genomes

Table 5e. Total number of novel repeats/domains in *Archaeoglobus fulgidus* DSM 4304 proteome.

S. No	Repeat / Domain name	Length of Repeat / Domain (amino acids)	Predicted Secondary structure	Number of Proteins identified from PSI-BLAST	Taxonomy of repeat occurrence
1	LIST (R)	45	2 β strands	2	*Archaeoglobus fulgidus* specific
2	SDL (D)	67	3 β strands	1	*Archaeoglobus fulgidus* specific
3	GSY (R)	41	2 β strands	1	*Archaeoglobus fulgidus* specific
4	CCE (D)	83	5 β strands	1	*Archaeoglobus fulgidus* specific
5	EES (D)	93	5 β strands	1	*Archaeoglobus fulgidus* specific
6	KEE (R)	32	3 α helices	1	*Archaeoglobus fulgidus* specific
7	DGVL (D)	55	ααββ	1	*Archaeoglobus fulgidus* specific
8	FQSP (R)	25	1β strand	1	*Archaeoglobus fulgidus* specific
9	CPAGCE (D)	74	coils	1	*Archaeoglobus fulgidus* specific
10	LAxY (D)	87	βββαβββ	1	*Archaeoglobus fulgidus* specific
11	FxP (D)	137	12 β strands	2	*Archaeoglobus fulgidus* and *Methanocaldococcus jannaschii*

Table 5f. Total number of novel repeats/domains in novel repeats/domains in *Halobacterium salinarium* NRC-1 proteome.

S. No	Repeat / Domain name	Length of Repeat / Domain (amino acids)	Predicted Secondary structure	Number of Proteins identified from PSI-BLAST	Taxonomy of repeat occurrence
1	GxW (D)	62	2 α helices	2	*Halobacterium salinarium & Haloarcula marismortui*
2	GxV (D)	64	5 β strands	1	*Halobacterium salinarium* specific
3	SCT (D)	55	2 β strands	1	*Halobacterium salinarium* specific
4	RxG (R)	37	3 β strands	2	*Halobacterium salinarium & Haloarcula marismortui*
5	LxT (D)	66	5 β strands	2	*Halobacterium salinarium & Haloarcula marismortui*
6	LEP (D)	120	βαββββββββ	2	*Halobacterium salinarium & Haloarcula marismortui*
7	HExxE (D)	106	6 α helices	2	*Halobacterium salinarium & Haloarcula marismortui*
8	PGE (D)	58	4 β strands	1	*Halobacterium salinarium* specific
9	VxA (D)	87	6 β strands	1	*Halobacterium salinarium* specific

Table 5g. Total number of novel repeats/domains in *Methanobacterium thermoautotrophicum* str. Delta H proteome.

S. No	Repeat / Domain name	Length of Repeat / Domain (aminoacids)	Predicted Secondary structure	Number of Proteins identified from PSI-BLAST	Taxonomy of repeat occurrence
1	VxV (D)	66	βαβ	3	*Methanothermobacter thermautotrophicus* and other Archaeal genomes
2	CREC (D)	115	3 α helices	1	*Methanothermobacter thermautotrophicus* specific
3	CPG (D)	187	αβββββααα	15	*Methanothermobacter thermautotrophicus* and other Archaeal genomes
4	TPG (D)	148	βββββββαα	14	*Methanothermobacter thermautotrophicus* and other Archaeal genomes
5	RxP (R)	48	1 β strand	1	*Methanothermobacter thermautotrophicus* specific
6	YTxP (R)	45	5 β strands	1	*Methanothermobacter thermautotrophicus* specific

Table 5h. Total number of novel repeats/domains in *Methanocaldococcus jannaschii* DSM 2661 proteome.

S. No	Repeat / Domain name	Length of Repeat / Domain (amino acids)	Predicted Secondary structure	Number of Proteins identified from PSI-BLAST	Taxonomy of repeat occurrence
1	GYI (D)	91	ααββ	1	*Methanocaldococcus jannaschii* specific
2	IPDY (D)	58	3 β strands	1	*Methanocaldococcus jannaschii* specific
3	CGA (R)	31	loops	1	*Methanocaldococcus jannaschii* specific
4	IxE (D)	90	4 α helices	1	*Methanocaldococcus jannaschii* specific
5	CG (R)	27	2 β strands	4	*Methanocaldococcus jannaschii* and other Archaeal & Bacterial genomes
6	FYD (D)	185	βαβββββββ	1	*Methanocaldococcus jannaschii* specific
7	TLY (D)	66	αααβα	1	*Methanocaldococcus jannaschii* specific

Table 5i. Total number of novel repeats/domains in *Methanopyrus kandleri* AV19 proteome.

S. No	Repeat / Domain name	Length of Repeat / Domain (amino acids)	Predicted Secondary structure	Number of Proteins identified from PSI-BLAST	Taxonomy of repeat occurrence
1	DWCA (D)	100	6 β strands	2	*Methanopyrus kandleri* specific
2	DYG (D)	213	10 β strands	2	*Methanopyrus kandleri* specific

Table 5j. Total number of novel repeats/domains in *Methanosarcina acetivorans* str. C2A proteome.

S. No	Repeat / Domain name	Length of Repeat / Domain (amino acids)	Predicted Secondary structure	Number of Proteins identified from PSI-BLAST	Taxonomy of repeat occurrence
1	PLM (D)	58	3 α helices	3	*M. acetivorans* and other Archaeal genomes
2	KLK (R)	24	1 α helix	1	*M. acetivorans* specific
3	LSW (D)	135	4 β strands	2	*M. acetivorans* and Bacterial genome
4	STS (D)	85	ααββα	1	*M. acetivorans* specific
5	SIV (R)	28	3 α heices	3	*M. acetivorans* and other Archaeal genomes
6	GLW (D)	232	αβββαββ	2	*M. acetivorans* and other Archaeal genomes
7	DDR (R)	24	1α helix	1	*M. acetivorans* specific
8	TQN (R)	17	1α helix	1	*M. acetivorans* specific
9	PxL (R)	36	2 α helices	4	*M. acetivorans* and other Archaeal genomes
10	ELI (R)	43	βα	1	*M. acetivorans* specific
11	KPE (R)	36	loops	3	*M. acetivorans* and other Archaeal genomes
12	GGY (D)	103	5 β strands	6	*M. acetivorans* and other Archaeal genomes
13	LVC (R)	49	ββαβ	4	*M. acetivorans* and other Archaeal genomes
14	NLE (R)	46	4 α helices	4	*M. acetivorans* and other Archaeal genomes
15	GE (R)	42	2 β strands	1	*M. acetivorans* specific
16	NLG (R)	36	3 α helices	1	*M. acetivorans* specific
17	GxD (D)	92	4 β strands	1	*M. acetivorans* specific
18	FNP (R)	23	1 α helix	1	*M. acetivorans* specific
19	AIK (D)	133	7 α helices	2	*M. acetivorans* and other Archaeal genomes
20	WVP (R)	26	2 β strands	1	*M. acetivorans* specific
21	YP (D)	57	4 α helices	68	Archaeal and Bacterial genomes

Table 5k. Total number of novel domains in *Pyrococcus abyssi* GE5 proteome.

S. No	Repeat / Domain name	Length of Repeat / Domain (amino acids)	Predicted Secondary structure	Number of Proteins identified from PSI-BLAST	Taxonomy of repeat occurrence
1	GYS (D)	146	βαβαβαα	4	*Pyrococcus abyssi* and other Archaeal genomes
2	GxF (D)	59	5 β strands	3	*Pyrococcus abyssi* and other Archaeal genomes
3	VTI (D)	56	6 β strands	3	*Pyrococcus abyssi* and other Archaeal genomes
4	NxG (D)	100	5 β strands	7	*Pyrococcus abyssi* and other Archaeal genomes

Table 5l. Total number of novel domains in *Thermoplasma acidophilum* DSM 1728 proteome.

S. No	Repeat / Domain name	Length of Repeat / Domain (amino acids)	Predicted Secondary structure	Number of Proteins identified from PSI-BLAST	Taxonomy of repeat occurrence
1	GLP (D)	77	7 β strands	2	*Thermoplasma acidophilum* and *Thermoplasma volcanium*
2	GxY (D)	69	4 β strands	1	*Thermoplasma acidophilum* specific
3	IK (D)	91	αααββ	4	*Thermoplasma acidophilum* and other Archaeal genomes
4	TxN (D)	167	6 β strands	4	*Thermoplasma acidophilum* and other Archaeal genomes

Table 5m. Total number of novel repeats/domains in *Picrophilus torridus* DSM 9790 proteome.

S. No	Repeat / Domain name	Length of Repeat / Domain (amino acids)	Predicted Secondary structure	Number of Proteins identified from PSI-BLAST	Taxonomy of repeat occurrence
1	YxxxG (R)	30	3 β strands	1	*Picrophilus torridus* specific
2	IYQ (R)	51	βββα	1	*Picrophilus torridus* specific
3	YKL (R)	51	5 β strands	1	*Picrophilus torridus* specific
4	NNT (R)	44	3 β strands	5	*Picrophilus torridus* and other Archaeal genomes
5	AW (R)	40	2 α helices	2	*Picrophilus torridus* and *Ferroplasma acidarmanus*
6	YN (R)	42	4 β strands	1	*Picrophilus torridus* specific

Table 5n. Total number of novel repeats/domains in *Nanoarchaeum equitans* Kin4-M proteome.

S. No	Repeat / Domain name	Length of Repeat / Domain (amino acids)	Predicted Secondary structure	Number of Proteins identified from PSI-BLAST	Taxonomy of repeat occurrence
1	DxLN (D)	55	βααα	1	*Nanoarchaeum equitans* specific
2	YGGK (R)	33	1 β strand	1	*Nanoarchaeum equitans* specific

The tables indicate the name of the novel repeats and domains, their length, predicted secondary structure, number of proteins identified from BLAST and the Taxonomy of these repeats and domains occurrence in other archaeal and bacterial genomes. (R) represents repeat and (D) represents domain

Figure 5.1: Phylogenetic tree of life based on differences in rRNA showing the separation of bacteria, archaea and eukaryotes.

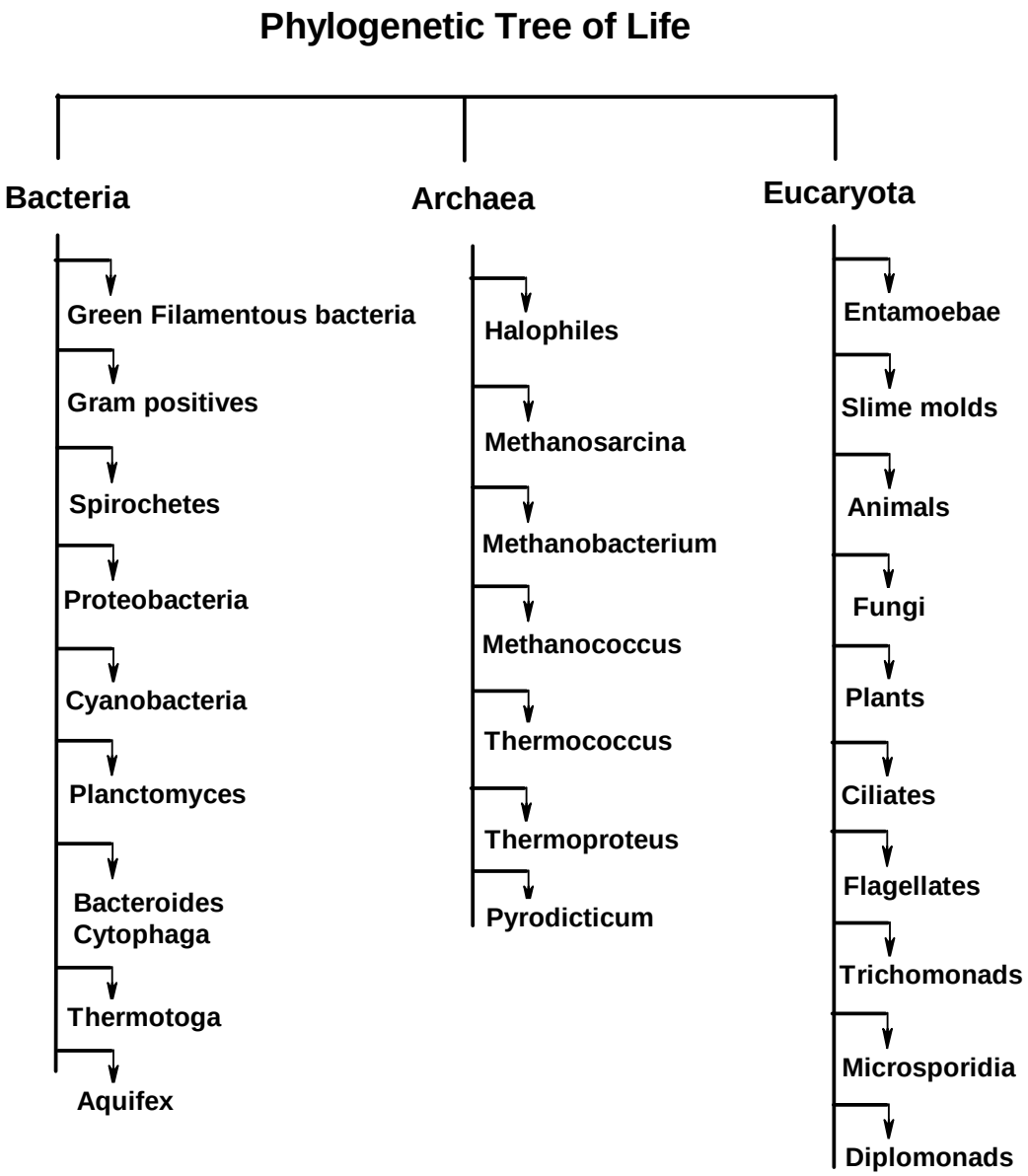
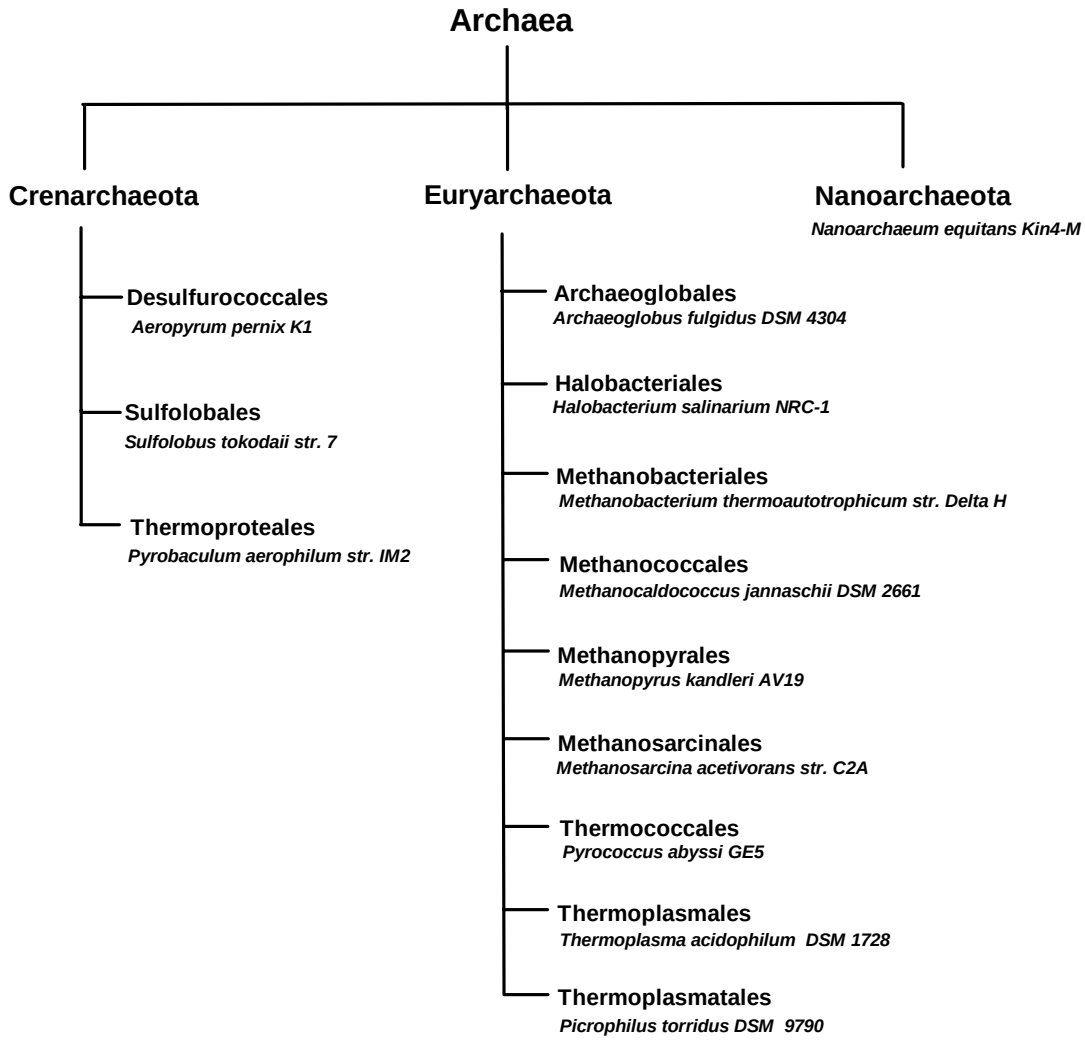


Figure 5.2: Phylogeny of archaeans based on molecular sequences in their DNA along with the repeat analyzed organisms.



5.4 Conclusions

1. A systematic *in silico* analysis of 13 representative archaeal genomes according to the phylogeny using computational tools, identified 56 novel domains and 38 repeats.

2. We observed that the 100 amino acid residues GQP domain of *S. tokodaii* str. 7 belongs to COG1449 (Cluster of Orthologues) and the domain is predicted to function as sugar transporter permease protein.

3. The 108 amino acid residues NDFA domain of *P. aerophilum* str. IM2 belongs to COG0438M and is predicted to function as trehalose-6-phosphate synthase.

4. The 83 amino acid residues CCE domain of *A. fulgidus* DSM 4304 has been described as a cell surface protein and we propose that these are cell surface protein specific domains.

5. From the repeats and domains present in these representative archaeal genomes, we infer that *N. equitans* NRC-1 is a minimalist archaea. The exchange of genes between archaeal and bacterial genomes is maximal in *M. acetivorans* C2A.

6. The number of orphan proteins comprising repeats and domains is high indicating the extensive evolution of these genomes. This is required for their adaptation to extreme living conditions such as high temperature, pressure and pH.

5.5 References

Adindla, S., Inampudi, K. K., Guruprasad, K. & Guruprasad, L. (2004). Identification and analysis of novel tandem repeats in the cell surface proteins of archaeal and bacterial genomes using computational tools. *Comparative and Functional Genomics,* **5,** 2-16.

Burns, D. G., Camakaris, H. M., Janssen, P. H. & Dyall-Smith, M. L. (2004). Cultivation of Walsby's square haloarchaeon. *FEMS Microbiol. Lett.* **238,** 469-473.

Danson, M: Central metabolism of the Archaea. In *The Biochemistry of Archaea (Archaeabacteria).* Edited by Kates M. New York: Elsevier Science Publishers; 1993:1-24.

DeLong, E. (1992). Archaea in coastal marine environments. *Proc. Natl. Acad. Sci. USA,* **89,** 5685-5689.

DeLong, E. F. & Pace, N. R. (2001). Environmental diversity of bacteria and archaea. *Syst. Biol*. **50,** 470-478.

Fuhrman, J., McCallum, K. & Davis, A. (1992). Novel major archaebacterial group from marine plankton. *Nature,* **356,** 148-149.

Gaasterland, T. (1999). Archaeal genomics. *Current Opinion in Microbiology,* **2,** 542-547.

Gao, B. & Gupta, R. S. (2007). Phylogenetic analysis of proteins that are distinctive of Archaea and its main subgroups and the origin of methanogenesis. *BMC Genomics*, **8,** 86.

Huber, Harald, Hohn, Michael. J., Rachel, Reinhard, Fuchs, *et al.* (2002). A new phylum of Archaea represented by a nanosized hyperthermophilic symbiont. *Nature,* **417,** 63-67.

Kates, M. (1992). Archaebacterial lipids: structure, biosynthesis and function. *Biochemical Society Symposium*, **58,** 51-72.

Valentine, D. L. (2007). Adaptations to energy stress dictate the ecology and evolution of the Archaea. *Nat. Rev. Microbiol*. **5,** 316-323.

Woese, C. & Fox, G. (1977). Phylogenetic structure of the prokaryotic domain: the primary kingdoms. *Proc. Natl. Acad. Sci. USA,* **74,** 5088–5090.

Woese C, Kandler, O. & Wheelis, M. (1990). Towards a natural system of organisms: proposal for the domains archaea, bacteria, and eucarya. *Proc. Natl. Acad. Sci. USA*, **87,** 4576-4579.

Woese, C: The Archaea: their history and significance. In *The Biochemistry of Archaea (Archaeabacteria).* Edited by Kates M. New York: Elsevier Science Publishers; 1993:v-xxxix.

CHAPTER 6

Identification and Analysis of Novel Amino Acid Sequence Repeats and Domains in Human Proteome

6.1 Introduction

The Human Genome Project (HGP) was launched in 1990 with the goal of obtaining a highly accurate sequence of the vast majority of the euchromatic portion of the human genome. The draft sequence of the human genome provides a huge challenge of how to interpret its biological function (I.H.G.S. Consortium 2001; Venter *et al*. 2001). A predominant part of the human genome consists of repetitive sequences of various types encompassing large segmental duplications, interspersed transposon derived repeats and tandem repeats (I.H.G.S. Consortium 2001). Often some of these repeats occur within genes and even within their coding sequences and perform regulatory functions. Also the repeats increase the likelihood of deleterious mutations in their host genes, thus increasing the risk of disease (Jasinska *et al*., 2004).

Amino acid tandem repeats, also known as homopolymeric tracts, is a very common feature of eukaryotic proteins (Green & Wang, 1994). They are present in nearly one-fifth of human gene products (Karlin *et al*., 2002, Alba & Guigo, 2004). Human proteins contain more amino acid repeats than rodent proteins and the trinucleotide repeats are also more abundant in human coding sequences (Alba & Guigo, 2004). Tandem repeats show a high degree of repeat unit length polymorphism, lie outside well defined structural/functional domains (Huntley & Golding, 2002) and tend to occur in sequences which are poorly conserved in evolution (Nishizawa *et al*., 1999). They are often embedded in low-complexity regions or simple sequences, which also include interrupted, non-tandem repeats. The low degree of conservation, or high turnover, of repeats may be related to a low degree of purifying selection (Hancock *et al*., 2001) and to the effect of trinucleotide slippage on expanding or shortening the repeats (Levinson & Gutman, 1987). At the genomic level, slippage of short DNA motifs (1-6 units) results in the formation of microsatellites, the length distribution of which can be modeled as a balance

between two evolutionary forces: slippage and point mutation (Kruglyak *et al*., 1998). Slippage can also have pathogenic effects: the uncontrolled expansion of trinucleotide repeats within human coding sequences is associated with several neurodegenerative disorders. Examples are Huntington's disease and dentatorubro-pallidolusyan atrophy, both associated with abnormally long expansions of CAG runs encoding poly-glutamine tracts (Wells, 1996, Gatchel & Zoghbi, 2005).

The high polymorphism and wide distribution of amino acid repeats may imply that in many cases they are functionally neutral. However, there is increasing biochemical evidence that in particular proteins some repeat types, such as glutamine, alanine, proline and glycine runs, can modulate protein-protein interactions and/or regulate transcription (Mitchell & Tjian 1989; Emili *et al.,* 1994; Gerber *et al.,* 1994; Perutz 1994; Imafuku *et al.,* 1998; Xiao & Jeang 1998; Wilkins & Lis 1999). In addition, tandem amino acid repeats do not appear in proteins in a random fashion; on the contrary, a significant association of different types of repeats with transcription factors and developmental proteins has been observed (Karlin & Burge 1996; Alba *et al.,* 1999a; Young *et al.,* 2000).

A domain is a structural or functional unit in a protein. A "domain" refers to a region of the protein comprising greater than 55 amino acid residues and does not contain internal sequence repeats. There are 1,865, 1,218, 1,183 and 973 domain types in human, fruit fly, nematode and yeast, respectively. Some proteins exhibit extensive domain repetition: in human, the largest number of domain types in a protein is nine, but the largest total number of domains in a protein is 130. Many human proteins have identical arrangements. There are also many human proteins that share more than one type of domain with fruit fly (slightly less frequently) with nematode and (much less frequently) with yeast proteins (Li *et al.,* 2001). Discovery of highly variable

amino acid tandem repeats can thus help discover new loci that may be particularly prone to suffer repeat expansions and become pathogenic.

Repeat structures in proteins have recently been found to play vital roles in various biological functions ranging from signal transduction, transcription regulation, to apoptosis, but are also recognized by their association with several human diseases. It is of paramount importance to identify the structures of the individual protein repeats lying within the human proteome and explore their protein interaction mechanisms to understand the complex biological processes and the human body in itself. Realizing the importance of amino acid repeats in the proteome and in human disorders, we undertook a study to identify and analyze the novel amino acid sequence repeats that are not present in any of the known databases and that are not reported so far with the available draft sequence of human genome.

In this work, we have identified 7 domains and 18 repeats using TRUST (Szklarczyk & Heringa, 2004) that have not been reported before in human proteome. Lists of the total proteins containing these novel repeats and domains are shown in Table 6a (novel domains) and Table 6b (novel repeats). These tables indicate the names of the novel repeats and domains, their length, predicted secondary structure and their order, number of proteins identified from PSI-BLAST and the taxonomy of the organisms in other genomes. The secondary structural elements aligned well in the multiple sequence alignment.

Lists of the proteins containing these novel domains are shown in Tables 6.1a to 6.1g and novel repeats are shown in Tables 6.2a to 6.2r. These tables indicate the protein identifiers (GENE or Swall_ID), the number of amino acid residues in the protein, a description of the protein and other well characterized repeats and domains present in the protein. Some sequences representing these repeats or domains share lower than 15% pair-wise sequence identity. However, these sequences retain the conserved sequence motifs and the positions of secondary structure elements in the multiple sequence alignment.

For all the proteins, the amino acid sequence corresponding to each representative repeats (Figures from 6.1a to 6.1g) and domains (Figures from 6.2a to 6.2r) are shown in the multiple sequence alignments. Conservation of the position of secondary structural elements is indicated from the multiple sequence alignment. We discuss each of these novel repeats and domains in *Homo sapiens* as below.

6.2 Methods

Various methods used to carry out the repeat and domain identification have been discussed in Chapter 3.

6.3 Results and Discussion

The domains are as follows:

1. 58 amino acid residue GPA domain: The protein corresponding to the GENE_ID NP_001013707.1 comprising of 215 amino acid residues consists of 58 amino acid residue region as one copy. Further PSI-BLAST searches corresponding to the region (7-64) as query identified 4 proteins that are described as hypothetical (see Table 6.1a). The domain occurs as one copy in *H. sapiens* GENE_ID EAL23895.1 and *P. troglodytes* GENE_IDs XP_001174244.1 and XP_520604.2. The multiple sequence alignment identified GPA as conserved sequence motif. The pair-wise identities between sequences corresponding to GPA domain varied from 47-60%. The length of the proteins varied from 140 to 270 amino acid residues. The secondary structure is predicted to comprise 3 loops as shown in Figure 6.1a.

2. 61 amino acid residue RxH domain: The protein corresponding to the GENE_ID NP_835260.2 comprising of 2839 amino acid residues consists of 61 amino acid residue region as two copies. Further PSI-BLAST searches corresponding to the region (1915-1975) as query identified 15 proteins that are described as PDZ domain containing proteins (see Table 6.1b). The domain occurs as two copies in tandem in *H. sapiens, M. musculus, R. norvegicus* and *B. taurus* genomes and as one copy in *C. lupus familiaris*. The PDZ domain is a common structural domain of 80-90 amino acids found in the signaling proteins of bacteria, yeast, plants and animals (Ponting, 1997). The PDZ domain is a widespread protein module that has been recruited to serve multiple functions during the course of evolution. These domains are found in various proteins in humans, alone or in arrays and in particular, they play prominent roles in synapse formation in mammals (Kim & Sheng, 2004). The multiple sequence alignment identified RxH as conserved sequence motif (where x is any amino acid residue). The pair-wise identities between sequences corresponding to

RxH domain varied from 20-100%. The length of the proteins varied from 1290 to 2847 amino acid residues. The secondary structure is predicted to comprise of 2 coils as shown in Figure 6.1b.

3. 68 amino acid residue GLG domain: The protein corresponding to the GENE_ID XP_496331.2 comprising of 569 amino acid residues and consisting of 68 amino acid residue region as seven copies (see Table 6.1c). Further PSI-BLAST searches corresponding to the region (77-144) as query identified seven copies (3+4(tandem)) within the same protein and therefore this domain is *H. sapiens* specific. The multiple sequence alignment identified GLG as conserved sequence motif along with LSCS motif. The pair-wise identities between sequences corresponding to GLG domain varied between 47-98%. The secondary structure is predicted to comprise of 1 α helix as shown in Figure 6.1c. The GLG domain containing protein with GENE_ID XP_496331.2 has been described as myosin XV protein. Myosins are mechanoenzymes defined by their conserved NH_2-terminal head or motor domains which contain actin- and adenosine triphosphate (ATP)-binding sites followed by a variable number of light-chain binding (IQ) motifs in the neck or flexible region and a variable tail domain. Upon interaction with actin, myosins convert energy from ATP hydrolysis to mechanical force as they pull against or move along actin filaments (Mooseker & Cheney, 1995). Myosins are presumed to acquire their specialized functions via their tails, which are tethered to different macro-molecular structures that move relative to actin filaments (Cheney *et al*., 1993). The tails of myosin XV and myosin VIIa share several regions of amino acid identity (Liang *et al*., 1999). *Myo15* encodes an unconventional myosin (myosin XV) that is mutated in the shaker-2 (*sh2*) and shaker-2J (*sh2J*) mice, and *DFNB3*, a form of non-syndromic hearing loss in humans (Liang *et al*., 1999, Probst *et al*., 1998, Friedman *et al*., 2000, Wang *et al*., 1998, Liang *et al*., 1998). *Myo15* mutant mice are congenitally deaf and have vestibular defects associated with circling behavior (Liang *et al*., 1998).

4. 71 amino acid residue SAS domain: The protein corresponding to the GENE_ID XP_209234.5 comprising of 1468 amino acid residues and described as hypothetical protein consists of 71 amino acid residue region as one copy (see Table 6.1d). Further PSI-BLAST searches corresponding to the region (489-559) as query identified one copy in 4 proteins of *H. sapiens.* The protein corresponding to the GENE_ID CAI15880.1 is described as Chromosome 1open reading frame 167. Chromosome 1 open reading frame 10 (*C1orf10*) gene is a recently identified gene that encodes a protein characterized with the presence of an EF-hand calcium-binding motif similar to S100 proteins, a conserved repeated sequence with similarity to bacterial ice nucleation protein, and one transmembrane domain (Xu *et al*., 2000). *C1orf10* was originally identified by differential display polymerase chain reaction (PCR) as one of the down-regulated genes in esophageal carcinoma (Xu *et al*., 1999). Its expression was highly detected in esophageal mucosa and dramatically reduced or absent in esophageal cancer cell lines and primary esophageal tumor tissues (Xu *et al*., 2000). The multiple sequence alignment identified SAS as conserved sequence motif. The pair-wise identities between sequences corresponding to SAS domain varied from 98-100%. The length of the proteins varied from 504 to 1468 amino acid residues. The secondary structure is predicted to comprise of 3 α helices as shown in Figure 6.1d. The SAS domain is *H. sapiens* specific domain.

5. 73 amino acid residue WKRK domain: The protein corresponding to the GENE_ID NP_001092905 comprising of 337 amino acid residues consists of 73 amino acid residue region as two copies in tandem. Further PSI-BLAST searches corresponding to the region (14-86) as query identified 37 proteins that are described as Williams-Beuren syndrome critical region protein 19, isoform CRA_a proteins as well as hypothetical proteins (see Table 6.1e). The domain occurs in variable copy numbers of 1, 2 and 5 in *H. sapiens, M. mulatta* and *P. troglodytes.* The multiple sequence alignment identified WKRK as conserved

sequence motif along with many other motifs such as APEPEE, LCGLKMK and LPE. The pair-wise identities between sequences corresponding to WKRK domain varied from 68-98%. The length of the proteins varied from 107 to 549 amino acid residues. The secondary structure is predicted to comprise of 3 α helices and 1 β as shown in Figure 6.1e. Williams-Beuren syndrome (WBS; OMIM 194050) is caused by heterozygous deletions of ~1.6 Mb of chromosomal sub-band 7q11.23. The deletions are rather uniform in size as they arise spontaneously by inter or intra-chromosomal crossover events within misaligned duplicated regions of high sequence identity that flank the typical deletion (Uta Francke, 1999).

6. 85 amino acid residue FSS domain: The protein corresponding to the GENE_ID NP_057047.3 comprising of 577 amino acid residues consists of 85 amino acid residue region as two copies. Further PSI-BLAST searches corresponding to the region (168-252) as query identified 14 proteins that are described as THAP domain containing proteins (see Table 6.1f). The domain occurs as two copies in *H. sapiens, B. taurus, C. lupus familiaris, M. mulatto, M. musculus, P. troglodytes, R. norvegicus* and as one copy in *H. sapiens* with GENE_ID AAH71896.1. The multiple sequence alignment identified FSS as conserved sequence motif along with other motifs such as SGACK and SLHSY. The pair-wise identities between sequences corresponding to FSS domain varied from 21-98%. The length of the proteins varied from 330 to 686 amino acid residues. THAP domain is a novel example of DNA-binding domain shared between cellular proteins and transposases from mobile genomic parasites (Roussigne *et al*, 2003) and we predict a similar function for the FSS domain. The secondary structure is predicted to comprise of 2 α helices as shown in Figure 6.1f.

7. 109 amino acid residue LLE domain: The protein corresponding to the GENE_ID NP_055685.2 comprising of 1239 amino acid residues consists of 109 amino acid residue region as two copies in tandem. Further PSI-BLAST searches corresponding to the region (237-345) as query identified 19 proteins that are described as Zinc finger and BTB domain containing proteins (see Table 6.1g). The domain occurs as two copies in *H. sapiens, B. taurus, C. lupus familiaris, M. mulatto, M. musculus, P. troglodytes, R. norvegicus* and as one copy in *H. sapiens* with GENE_ID AAI14608.1, *P. troglodytes* GENE_ID XP_001165023.1, *B. taurus* GENE_ID XP_001253042.1 and *M. musculus* GENE_ID XP_919018.2. The multiple sequence alignment identified LLE as conserved sequence motif. The pair-wise identities between sequences corresponding to LLE domain varied from 4-98%. The length of the proteins varied from 453 to 1258 amino acid residues. The secondary structure is predicted to comprise of 5 α helices as shown in Figure 6.1g. Zinc finger domains fall into more than twenty subclasses, based on their fold and zinc ligation topology and different members can mediate interactions with DNA, RNA, proteins and other molecules (Gamsjaeger *et al*., 2007).

The RxH, WKRK, FSS and LLE domains are present in *Homo sapiens* and other eukaryotic genomes, GPA domain is present in *Homo sapiens* and *Pan troglodytes* genomes. The GLG and SAS domains are *Homo sapiens* specific and are orphan proteins (see Table 6a).

The repeats are as follows:

1. 30 amino acid residue PGQY repeat: The protein corresponding to the GENE_ID XP_059954.3 comprising of 237 amino acid residues consists of 30 amino acid residue region as two copies. Further PSI-BLAST searches corresponding to the region (121-150) as query identified 5 proteins that are described as hypothetical, RP11-346E17.3 and C9orf57 proteins (see Table 6.2a). The repeat occurs as two copies in *H. sapiens* and *P. troglodytes* and as one copy in *B. taurus* with GENE_ID XP_001253313.1. The multiple sequence alignment identified PGQY as conserved sequence motif. The pair-wise identities between sequences corresponding to PGQY repeat varied from 43-96%. The length of the proteins varied from 127 to 237 amino acid residues. The secondary structure is predicted to comprise of 1 α helix as shown in Figure 6.2a.

2. 31 amino acid residue FYE repeat: The protein corresponding to the GENE_ID NP_115754.2 comprising of 647 amino acid residues consists of 31 amino acid residue region as two copies. Further PSI-BLAST searches corresponding to the region (222-252) as query identified 15 proteins that are described as Eukaryotic translation elongation factor 1 delta (see Table 6.2b). The repeat occurs as two copies in *Danio rerio* and as one copy in *H. sapiens, B. taurus, C. lupus familiaris, M. musculus, M. fascicularis, P. troglodytes, R. norvegicus* and *G. gallus*. The multiple sequence alignment identified FYE as conserved sequence motif along with AER motif. The pair-wise identities between sequences corresponding to FYE repeats varied from 25-90%. The length of the proteins varied from 550 to 679 amino acid residues. The secondary structure is predicted to comprise of 2 α helices as shown in Figure 6.2b.

3. 34 amino acid residue VHMM repeat: The protein corresponding to the GENE_ID NP_001072997.2 comprising of 387 amino acid residues consists of 34 amino acid residue region as three copies (1+2 tandem). Further PSI-

BLAST searches corresponding to the region (181-214) as query identified 10 proteins that are described as hypothetical, NY-REN-7 protein, LOC202134, KIAA0752 protein and AF155097_1 NY-REN-7 antigen proteins (see Table 6.2c). The repeat occurs in variable copy numbers of 1, 2 and 3 in *H. sapiens* and *P. troglodytes.* The multiple sequence alignment identified VHMM as conserved sequence motif. The pair-wise identities between sequences corresponding to VHMM repeats varied from 64-94%. The length of the proteins varied from 114 to 387 amino acid residues. The secondary structure is predicted to comprise of 2 α helices as shown in Figure 6.2c.

The protein corresponding to the GENE_ID NP_149020.1 comprising of 1395 amino acid residues consists of two types of repeats. 1. 34 amino acid residue TQG repeats and 2. 51 amino acid residue PES repeats.

4. 34 amino acid residue TQG repeat: The 34 amino acid repeat region occurs as two copies. Further PSI-BLAST searches corresponding to the region (478-511) as query identified 7 proteins that are described as Cyclin B3 proteins. The repeat occurs as two copies in *H. sapiens* and *P. troglodytes* and as one copy in *C. lupus familiaris* (see Table 6.2d). The multiple sequence alignment identified TQG as conserved sequence motif. The pair-wise identities between sequences corresponding to TQG repeats varied from 43-98%. The length of the proteins varied from 899 to 1395 amino acid residues. The secondary structure is predicted to comprise of coils as shown in Figure 6.2d.

5. 51 amino acid residue PES repeat: The 51 amino acid repeat region occurs as two copies. Further PSI-BLAST searches corresponding to the region (955-1005) as query identified 7 proteins that are described as Cyclin B3 proteins. The repeat occurs as two copies in *H. sapiens* and as one copy in *C. lupus familiaris* and *P. troglodytes* (see Table 6.2e). The multiple sequence alignment identified PES as conserved sequence motif. The pair-wise identities between sequences corresponding to PES repeats varied from 43-98%. The length of the proteins varied from 899 to 1395 amino acid residues. The secondary structure

is predicted to comprise of coils as shown in Figure 6.2e. Cyclin B3 is essential for fertility (Jacobs *et al*., 1998). *H. sapiens* cyclin B3 mRNA and protein are detected readily in developing germ cells in the human testis and not in any other tissue. Cyclin B3 is expressed in both males and females. In both sexes, it accumulates to its highest levels in zygotene cells and diminishes in pachytene cells. It either regulate events during the leptotene and zygotene, such as recombination or synapsis, or its turnover may be important for proper exit of cells from zygotene and their progression into pachytene (Nguyen *et al.,* 2002).

The protein corresponding to the GENE_ID XP_374705.3 consists of two types of repeats. 1. 34 amino acid residue HTQ repeats and 2. 38 amino acid residue PTT repeats.

6. 34 amino acid residue HTQ repeat: The 34 amino acid repeat region occurs as four copies (1+3 tandem). Further PSI-BLAST searches corresponding to the region (4-37) as query identified 3 proteins from *H. sapiens* that are described as Polycystic kidney disease 1 like 3 proteins (see Table 6.2f). The multiple sequence alignment identified HTQ as conserved sequence motif. The pair-wise identities between sequences corresponding to PES repeats varied from 73-97%. The length of the proteins varied from 683 to 687 amino acid residues. The secondary structure is predicted to comprise of coils as shown in Figure 6.2f. The HTQ repeats are *H. sapiens* specific.

7. 38 amino acid residue PTT repeat: The 38 amino acid repeat region occurs as six copies (3 (tandem) +3). Further PSI-BLAST searches corresponding to the region (280-317) as query identified 4 proteins from *H. sapiens* that are described as Polycystic kidney disease 1 like 3 proteins (see Table 6.2g). The repeat occurs as 3 copies in tandem in *H. sapiens* with GENE_ID NP_001078865. The multiple sequence alignment identified PTT as conserved sequence motif. The pair-wise identities between sequences corresponding to PTT repeats varied from 50-97%. The length of the proteins varied from 437 to

687 amino acid residues. The secondary structure is predicted to comprise of coils as shown in Figure 6.2g. The PTT repeats are *H. sapiens* specific. We observed that in *H. sapiens* one protein with GENE_ID EAL23926.1 with a length of 437 amino acid residues comprises the PTT repeats but the HTQ repeats are absent. Polycystic kidney disease (PKD) is a disease of the nephron, characterized by the formation of multiple renal tubular cysts, leading to endstage renal failure. The most common form is autosomal dominant PKD (ADPKD) and is caused by mutations in the *PKD1* gene in 85% of cases or in *PKD2* in 10-15% (Wilson, 2004). We, therefore, predict a similar function for HTQ and PTT repeats.

8. 34 amino acid residue FSQ repeat: The protein corresponding to the GENE_ID NP_008917.3 comprising of 778 amino acid residues consists of 34 amino acid residue region as three copies in tandem. Further PSI-BLAST searches corresponding to the region (70-103) as query identified 18 proteins that are described as Melanoma antigen family D 1 proteins (see Table 6.2h). The repeats occurs as three copies in tandem in *H. sapiens, B. taurus, C. lupus familiaris, M. fascicularis, M. musculus, R. norvegicus* and *S. scrofa* genomes. The multiple sequence alignment identified FSQ as conserved sequence motif. The pair-wise identities between sequences corresponding to FSQ repeats varied from 5-97%. The length of the proteins varied from 353 to 834 amino acid residues. The secondary structure is predicted to comprise of coils as shown in Figure 6.2h.

9. 36 amino acid residue PEG repeat: The protein corresponding to the GENE_ID NP_005453.2 comprising of 1142 amino acid residues consists of 36 amino acid residue region as three copies. Further PSI-BLAST searches corresponding to the region (12-48) as query identified 6 proteins that are described as Melanoma antigen family C 1 proteins (see Table 6.2i). The repeat occurs in variable copy numbers of 1, 2 and 3 in *H. sapiens* and therefore said to be *H. sapiens* specific. The multiple sequence alignment identified PEG as

conserved sequence motif. The pair-wise identities between sequences corresponding to PEG repeats varied from 59-97%. The length of the proteins varied from 118 to 1142 amino acid residues. The secondary structure is predicted to comprise of coils as shown in Figure 6.2i.

10. 42 amino acid residue SSC repeat: The protein corresponding to the GENE_ID XP_001127353.1 comprising of 299 amino acid residues consists of 42 amino acid residue region as one copy. Further PSI-BLAST searches corresponding to the region (151-192) as query identified 14 proteins that are described as hypothetical, Isoform CRA_c and Isoform CRA_a proteins (see Table 6.2j). The repeat occurs in variable copy numbers of 1, 2 and 3 copies in *H. sapiens, M. mulatto* and *P. troglodytes* genomes. The multiple sequence alignment identified SSC as conserved sequence motif. The pair-wise identities between sequences corresponding to FSQ repeats varied from 9-97%. The length of the proteins varied from 159 to 420 amino acid residues. The secondary structure is predicted to comprise of 1 α helix as shown in Figure 6.2j.

11. 42 amino acid residue YCL repeat: The protein corresponding to the GENE_ID NP_060880.3 comprising of 748 amino acid residues consists of 42 amino acid residue region as two copies. Further PSI-BLAST searches corresponding to the region (411-452) as query identified 17 proteins that are described as hypothetical proteins (see Table 6.2k). The repeat occurs in 1 or 2 copy numbers in *H. sapiens, B. taurus, M. fascicularis, M. mulatta, M. musculus, R. norvegicus,* and *P. troglodytes* genomes. The multiple sequence alignment identified YCL as conserved sequence motif. The pair-wise identities between sequences corresponding to YCL repeats varied from 21-97%. The length of the proteins varied from 388 to 858 amino acid residues. The secondary structure is predicted to comprise of 1 α helix and 1 β strand as shown in Figure 6.2k.

The protein corresponding to the GENE_ID XP_374142.2 comprising of 1015 amino acid residues consists of two types of repeats. 1. 43 amino acid residue VSR repeats and 2. 43 amino acid residue ALPG repeats.

12. 43 amino acid residue VSR repeat: The 43 amino acid residue region occurs as one copy. Further PSI-BLAST searches corresponding to the region (408-450) as query identified 6 proteins that are described as hypothetical proteins (see Table 6.2l). The repeat occurs as one copy in *H. sapiens, M. fascicularis, M. mulatta, M. musculus, R. norvegicus*, and *E. caballus* genomes. The multiple sequence alignment identified VSR as conserved sequence motif. The pair-wise identities between sequences corresponding to VSR repeats varied from 44-90%. The length of the proteins varied from 1006 to 1071 amino acid residues. The secondary structure is predicted to comprise of 1 α helix as shown in Figure 6.2l.

13. 54 amino acid residue ALPG repeat: The 54 amino acid residue region occurs as one copy. Further PSI-BLAST searches corresponding to the region (5-58) as query identified 6 proteins that are described as hypothetical proteins (see Table 6.2m). The repeat occurs as one copy in *H. sapiens, M. fascicularis, M. mulatta, M. musculus, R. norvegicus* and *E. caballus* genomes. The multiple sequence alignment identified ALPG as conserved sequence motif. The pair-wise identities between sequences corresponding to ALPG repeats varied from 77-98%. The length of the proteins varied from 1006 to 1071 amino acid residues. The secondary structure is predicted to comprise of loops as shown in Figure 6.2m.

14. 43 amino acid residue SVT repeat: The protein corresponding to the GENE_ID XP_499019.2 comprising of 376 amino acid residues and described as hypothetical protein consists of 43 amino acid residue region as six copies (4 tandem +2) (see Table 6.2n). Further PSI-BLAST searches corresponding to the region (39-81) as query identified six copies within the same protein and therefore it is *H. sapiens* specific. The multiple sequence alignment identified

SVT as conserved sequence motif. The pair-wise identities between sequences corresponding to SVT repeats varied from 81-97%. The secondary structure is predicted to comprise of coils as shown in Figure 6.2n.

15. 49 amino acid residue CDxD repeat: The protein corresponding to the GENE_ID NP_003226.4 comprising of 2768 amino acid residues consists of 49 amino acid residue region as two copies. Further PSI-BLAST searches corresponding to the region (1708-1753) as query identified 18 proteins that are described as Thyroglobulin precursor proteins (see Table 6.2o). The repeat occurs in two copies in *H. sapiens, B. taurus R. norvegicus* and as one copy in *C. lupus familiaris* GENE_ID NP_001041569.1, *T. nigroviridis* GENE_ID CAF89701.1 and *R. norvegicus* GENE_ID CAA26183.1 genomes. The multiple sequence alignment identified CDxD as conserved sequence motif. The pair-wise identities between sequences corresponding to CDxD repeats varied from 18-97%. The length of the proteins varied from 967 to 2769 amino acid residues. The secondary structure is predicted to comprise of 1 α helix and 3 β strands as shown in Figure 6.2o. Thyroglobulin is the primary synthetic product of the thyroid and the macromolecular precursor of thyroid hormones (Suzuki *et al.,* 1999).

16. 50 amino acid residue GGF repeat: The protein corresponding to the GENE_ID XP_497341.3 comprising of 7328 amino acid residues consists of 50 amino acid residue region as three copies in tandem. Further PSI-BLAST searches corresponding (116-165) as query identified 2 proteins that are described as mucin 19 proteins from *H. sapiens* and therefore it is *H. sapiens* specific (see Table 6.2p). The multiple sequence alignment identified GGF as conserved sequence motif. The pair-wise identities between sequences corresponding to GGF repeats varied from 46-100%. The length of the proteins varied from 4516 to 7329 amino acid residues. The secondary structure is predicted to comprise of coils as shown in Figure 6.2p.

17. 52 amino acid residue NYS repeat: The protein corresponding to the GENE_ID NP_048536.2 comprising of 1299 amino acid residues and described as SWI/SNF chromatin remodeling complex subunit OSA2 protein consists of 52 amino acid residue region as four copies in tandem (see Table 6.2q). Further PSI-BLAST searches corresponding (151-203) as query identified four copies within the same protein and therefore it is *H. sapiens* specific. The multiple sequence alignment identified NYS as conserved sequence motif. The pair-wise identities between sequences corresponding to NVT repeats varied from 57-92%. The secondary structure is predicted to comprise of coils as shown in Figure 6.2q.

18. 52 amino acid residue RPE repeat: The protein corresponding to the GENE_ID NP_835260.2 comprising of 2839 amino acid residues consists of 52 amino acid residue region as one copy. Further PSI-BLAST searches corresponding to the region (1234-1286) as query identified 8 proteins that are described as PDZ domain containing proteins (see Table 6.2r). The repeat occurs as one copy in *H. sapiens, B. taurus C. lupus familiaris* and *P. troglodytes* genomes. The multiple sequence alignment identified RPE as conserved sequence motif along with SVR and RSP motifs. The pair-wise identities between sequences corresponding to RPE repeats varied from 45-98%. The length of the proteins varied from 2443 to 2847 amino acid residues. The secondary structure is predicted to comprise of 1 β strand as shown in Figure 6.2r.

The PGQY, FYE, VHMM, TQG, PES, FSQ, SSC, YCL, VSR, ALPG, CDxD and RPE repeats are present in *Homo sapiens* and other eukaryotic genomes. The HTQ, PTT, PEG, SVT, GGF and NYS repeats are *Homo sapiens* specific and are orphan proteins (see Table 6b).

Table 6a. Total Number of Novel Amino Acid Sequence Domains in Human Proteome.

S. No.	Domain name	Number of PSI-BLAST identified proteins	Predicted Secondary structure	Length of Domains (amino acids)	Taxonomy of Domain occurrence
1	GPA domain	4	3 loops	58-aa	*Homo sapiens* and *Pan troglodytes*
2	RxH domain	15	coils	61-aa	*Homo sapiens* and other eukaryotic genomes
3	GLG domain	1	1 α helix	68-aa	*Homo sapiens* specific
4	SAS domain	4	3 α helices	71-aa	*Homo sapiens* specific
5	WKRK domain	37	αβαα	73-aa	*Homo sapiens* and other eukaryotic genomes
6	FSS domain	14	2 α helices	85-aa	*Homo sapiens* and other eukaryotic genomes
7	LLE domain	19	5 α helices	109-aa	*Homo sapiens* and other eukaryotic genomes

The table (6a and 6b) indicate the name of the novel domains and novel repeats, their length, predicted secondary structure, number of proteins identified from BLAST and the Taxonomy of the repeat and domains occurrence in other eukaryotic genomes. (R) represents repeat and (D) represents domain.

Table 6b. Total Number of Novel Amino Acid Sequence Repeats in Human Proteome.

S. No.	Repeats name	Number of PSI-BLAST identified proteins	Predicted Secondary structure	Length of Repeats (amino acids)	Taxonomy of Repeat occurrence
1	PGQY repeat	5	1 α helix	30-aa	*Homo sapiens* and other eukaryotic genomes
2	FYE repeat	16	2 α helices	31-aa	*Homo sapiens* and other eukaryotic genomes
3	VHMM repeat	10	2 α helices	34-aa	*Homo sapiens* and other eukaryotic genomes
4	TQG repeat	7	coils	34-aa	*Homo sapiens* and other eukaryotic genomes
5	PES repeat	7	coils	51-aa	*Homo sapiens* and other eukaryotic genomes
6	HTQ repeat	3	coils	34-aa	*Homo sapiens* specific
7	PTT repeat	4	coils	38-aa	*Homo sapiens* specific
8	FSQ repeat	18	coils	34-aa	*Homo sapiens* and other eukaryotic genomes
9	PEG repeat	6	coils	36-aa	*Homo sapiens* specific
10	SSC repeat	14	1 α helix	42-aa	*Homo sapiens* and other eukaryotic genomes
11	YCL repeat	17	βα	42-aa	*Homo sapiens* and other eukaryotic genomes
12	VSR repeat	6	1 α helix	43-aa	*Homo sapiens* and other eukaryotic genomes
13	ALPG repeat	6	loops	54-aa	*Homo sapiens* and other eukaryotic genomes
14	SVT repeat	1	coils	43-aa	*Homo sapiens* specific
15	CDxD repeat	18	αβββ	49-aa	*Homo sapiens* and other eukaryotic genomes
16	GGF repeat	2	coils	50-aa	*Homo sapiens* specific
17	NYS repeat	1	coils	52-aa	*Homo sapiens* specific
18	RPE repeat	8	1β strand	53-aa	*Homo sapiens* and other eukaryotic genomes

Table 6.1a. List of proteins containing the 58 amino acid residue GPA domain.

GENE_ID (number of residues)	Organism	Description	Number of GPA domains
NP_001013707.1 (215)	*Homo sapiens* (E)	Hypothetical protein	1
EAL23895.1 (140)	*Homo sapiens* (E)	Hypothetical protein	1
XP_001174244.1 (155)	*Pan troglodytes* (E)	Hypothetical protein	1
XP_520604.2 (270)	*Pan troglodytes* (E)	Hypothetical protein	1

Table 6.1b. List of proteins containing the 61 amino acid residue RxH domain.

GENE_ID (number of residues)	Organism	Description	Number of RxH domains
NP_075229.1 (2766)	*Rattus norvegicus* (E)	PDZ domain containing 3 protein	2 (tandem)
XP_981908.1 (2802)	*Mus musculus* (E)	PDZ domain containing 3 protein	2 (tandem)
NP_001074533.1 (2796)	*Mus musculus* (E)	PDZ domain containing 3 protein	2 (tandem)
XP_912272.1 (2797)	*Mus musculus* (E)	PDZ domain containing 3 protein	2 (tandem)
XP_892828.1 (2744)	*Mus musculus* (E)	PDZ domain containing 3 protein	2 (tandem)
BAA20760.2 (2847)	*Homo sapiens* (E)	PDZ domain containing 3 protein	2 (tandem)
NP_835260.2 (2839)	*Homo sapiens* (E)	PDZ domain containing 2 protein	2 (tandem)
O15018 (2839)	*Homo sapiens* (E)	PDZ domain containing protein	2 (tandem)
AAK07661.1 (2641)	*Homo sapiens* (E)	PDZ domain containing protein	2 (tandem)
EAX10777.1 (2665)	*Homo sapiens* (E)	PDZ domain containing 2 protein	2 (tandem)
XP_526957.2 (2443)	*Pan troglodytes* (E)	PDZ domain containing 2 protein	2 (tandem)
AAI15887.1 (1290)	*Mus musculus* (E)	PDZ domain containing 2 protein	2 (tandem)
BAC65522.1 (1352)	*Mus musculus* (E)	PDZ domain containing 2 protein	2 (tandem)
XP_871254.2 (2803)	*Bos taurus* (E)	PDZ domain containing 2 protein	2 (tandem)
XP_536512.2 (2601)	*Canis lupus familiaris* (E)	PDZ domain containing 2 protein	1

Table 6.1c. Protein containing 68 amino acid residue GLG domain.

GENE_ID (number of residues)	Organism	Description	Number of GLG domains
XP_496331.2 (569)	*Homo sapiens* (E)	Similar to myosin XV	3+4 (tandem)

Table 6.1d. List of proteins containing the 71 amino acid residue SAS domain.

GENE_ID (number of residues)	Organism	Description	Number of SAS domains
CAI15880.1 (504)	*Homo sapiens* (E)	Chromosome 1open reading frame 167	1
XP_209234.5 (1468)	*Homo sapiens* (E)	Hypothetical protein	1
CAD38776.1 (509)	*Homo sapiens* (E)	Hypothetical protein	1
XP_943296.2 (1468)	*Homo sapiens* (E)	Hypothetical protein	1

Table 6.1e. List of proteins containing the 73 amino acid residue WKRK domain.

GENE_ID (number of residues)	Organism	Description	Number of WKRK domains
NP_001092905.1 (337)	*Homo sapiens* (E)	Hypothetical protein	2 (tandem)
EAW50235.1 (217)	*Homo sapiens* (E)	Similar to Williams-Beuren syndrome critical region protein 19, isoform CRA_a	1
XP_001128093.1 (338)	*Homo sapiens* (E)	Similar to Williams-Beuren syndrome critical region protein 19, isoform CRA_a	2 (tandem)
XP_001134912.1 (376)	*Pan troglodytes* (E)	Similar to Williams-Beuren syndrome critical region protein 19, isoform CRA_a	2 (tandem)
XP_001110275.1 (431)	*Macaca mulatta* (E)	Similar to Williams-Beuren syndrome critical region protein 19, isoform CRA_a	2 (tandem)
XP_001142986.1 (338)	*Pan troglodytes* (E)	Similar to Williams-Beuren syndrome critical region protein 19, isoform CRA_a	2 (tandem)
XP_001156328.1 (336)	*Pan troglodytes* (E)	Similar to Williams-Beuren syndrome critical region protein 19, isoform CRA_a	2 (tandem)
XP_001152570.1 (314)	*Pan troglodytes* (E)	Hypothetical protein isoform 2	1
XP_001150008.1 (266)	*Pan troglodytes* (E)	Hypothetical protein	1
XP_001149803.1 (265)	*Pan troglodytes* (E)	Hypothetical protein	1
XP_001152377.1 (265)	*Pan troglodytes* (E)	Hypothetical protein	1
XP_001156047.1 (107)	*Pan troglodytes* (E)	Similar to Williams-Beuren syndrome critical region protein 19	1
XP_940755.1 (265)	*Homo sapiens* (E)	Hypothetical protein	1
XP_496899.1 (265)	*Homo sapiens* (E)	Hypothetical protein	1
AAH45636.1 (257)	*Homo sapiens* (E)	LOC389517 protein	1
NP_778234.2 (336)	*Homo sapiens* (E)	Hypothetical protein	2 (tandem)
AAM62309.1 (336)	*Homo sapiens* (E)	Williams-Beuren syndrome critical region protein 19	2 (tandem)
CAB70665.1 (308)	*Homo sapiens* (E)	Hypothetical protein	1
XP_001135002.1 (107)	*Pan troglodytes* (E)	Similar to Williams-Beuren syndrome critical region protein 19	1
XP_001115801.1 (138)	*Macaca mulatta* (E)	Similar to speedy B, partial	1
XP_001152762.1 (147)	*Pan troglodytes* (E)	Hypothetical protein	1
XP_001130563.1 (399)	*Homo sapiens* (E)	Similar to Williams-Beuren syndrome chromosome	2 (tandem)

		region 19	
XP_001130493.1(368)	*Homo sapiens* (E)	Similar to Williams-Beuren syndrome chromosome region 19	2 (tandem)
XP_499348.2 (257)	*Homo sapiens* (E)	Hypothetical protein	1
EAW50237.1 (280)	*Homo sapiens* (E)	Similar to Williams-Beuren syndrome critical region protein 19, isoform CRA_c	1
XP_935532 (337)	*Homo sapiens* (E)	Similar to Williams-Beuren syndrome chromosome region 19 isoform 2	2 (tandem)
EAX07879.1 (337)	*Homo sapiens* (E)	hCG1815881	2 (tandem)
XP_499356.1 (337)	*Homo sapiens* (E)	Similar to Williams-Beuren syndrome chromosome region 19 isoform 1	2 (tandem)
XP_001127201.1 (337)	*Homo sapiens* (E)	Similar to Williams-Beuren syndrome chromosome region 19 isoform 1	2 (tandem)
XP_499314.2 (337)	*Homo sapiens* (E)	Similar to Williams-Beuren syndrome chromosome region 19 isoform 1	2 (tandem)
AAH56606 (549)	*Homo sapiens* (E)	MGC57359 protein	5 (tandem)
EAW94419.1 (290)	*Homo sapiens* (E)	hCG27838	2 (tandem)
AAI00975.1 (208)	*Homo sapiens* (E)	Williams-Beuren syndrome chromosome region 19 pseudogene	1
XP_371014.3 (237)	*Homo sapiens* (E)	Hypothetical protein	1
XP_001131690.1 (208)	*Homo sapiens* (E)	Similar to Williams-Beuren syndrome chromosome region 19 isoform 1	1
NP_001026789.1 (258)	*Homo sapiens* (E)	Hypothetical protein	1
NP_001004351.2 (172)	*Homo sapiens* (E)	Hypothetical protein	1

Table 6.1f. List of proteins containing the 85 amino acid residue FSS domain.

GENE_ID (number of residues)	Organism	Description	Number of FSS domains
NP_080196.3 (569)	*Mus musculus* (E)	THAP domain-containing protein	2
BAE24509.1 (569)	*Mus musculus* (E)	Unnamed protein product	2
EAW71276.1 (577)	*Homo sapiens* (E)	THAP domain-containing protein	2
NP_057047.3 (577)	*Homo sapiens* (E)	THAP domain-containing protein	2
Q8WY91 (577)	*Homo sapiens* (E)	THAP domain-containing protein	2
AAH57963.2 (436)	*Mus musculus* (E)	THAP domain-containing protein	2
XP_001093438.1 (442)	*Macaca mulatta* (E)	THAP domain-containing protein	2
XP_516210.2 (686)	*Pan troglodytes* (E)	THAP domain-containing protein	2
XP_001128859.1 (684)	*Homo sapiens* (E)	THAP domain-containing protein	2
NP_001005564.1 (569)	*Rattus norvegicus* (E)	THAP domain-containing protein	2
AAH66042.1 (379)	*Mus musculus* (E)	THAP domain-containing protein	2
NP_001033758.1 (570)	*Bos taurus* (E)	THAP domain-containing protein	2
XP_543333.2 (632)	*Canis lupus familiaris* (E)	THAP domain-protein	2
AAH71896.1 (330)	*Homo sapiens* (E)	THAP domain-containing protein	1

Table 6.1g. List of proteins containing the 109 amino acid residue LLE domain.

GENE_ID (number of residues)	Organism	Description	Number of LLE domains
XP_001164879.1 (1183)	*Pan troglodytes* (E)	Zinc finger and BTB domain containing 40 isoform 1	2 (tandem)
BAA32323.2 (1253)	*Homo sapiens* (E)	KIAA0478 protein	2 (tandem)
XP_001164989.1 (1192)	*Pan troglodytes* (E)	Zinc finger and BTB domain containing 40 isoform 3	2 (tandem)
Q9NUA8 (1239)	*Homo sapiens* (E)	ZBT40_HUMAN Zinc finger and BTB domain-containing protein 40	2 (tandem)
NP_055685.2 (1239)	*Homo sapiens* (E)	Zinc finger and BTB domain containing 40	2 (tandem)
XP_001164955.1 (1239)	*Pan troglodytes* (E)	Zinc finger and BTB domain containing 40 isoform 2	2 (tandem)
XP_001101017.1 (1147)	*Macaca mulatta* (E)	Similar to zinc finger and BTB domain containing 40 isoform 1	2 (tandem)
XP_001101280.1 (1192)	*Macaca mulatta* (E)	Similar to zinc finger and BTB domain containing 40 isoform 3	2 (tandem)
XP_001101193.1 (1239)	*Macaca mulatta* (E)	Similar to zinc finger and BTB domain containing 40 isoform 2	2 (tandem)
CAI22041.1 (453)	*Homo sapiens* (E)	Zinc finger and BTB domain containing 40	2 (tandem)
XP_544510.2 (1243)	*Canis lupus familiaris* (E)	Similar to zinc finger and BTB domain containing 40	2 (tandem)
XP_614579.2 (1232)	*Bos taurus* (E)	Similar to zinc finger and BTB domain containing 40 isoform 1	2 (tandem)
AAI14608.1 (1127)	*Homo sapiens* (E)	ZBTB40 protein	1
XP_001165023.1 (1127)	*Pan troglodytes* (E)	Zinc finger and BTB domain containing 40 isoform 4	1
NP_937891.1 (1258)	*Mus musculus* (E)	Zinc finger and BTB domain containing 40	2 (tandem)
BAD90181.1 (1234)	*Mus musculus* (E)	mKIAA0478 protein	2 (tandem)
XP_001253042.1 (1121)	*Bos taurus* (E)	Similar to zinc finger and BTB domain containing 40	1
XP_342954.3 (1255)	*Rattus norvegicus* (E)	Similar to zinc finger and BTB domain containing 40	2 (tandem)
XP_919018.2 (998)	*Mus musculus* (E)	Similar to zinc finger and BTB domain containing 40	1

The proteins are represented by their corresponding GENE_ID along with the number of amino acid residues indicated in brackets in the first column. The organism and corresponding phylogeny are indicated in the second column; 'E' represents Eukaryota. The third column contains the description of the proteins containing the domains identified in the present work. The fourth column represents exclusively the total number of novel domains identified in this work.

Table 6.2a. List of proteins containing the 30 amino acid residue PGQY repeat.

GENE_ID (number of residues)	Organism	Description	Number of PGQY repeats
XP_059954.3 (237)	*Homo sapiens* (E)	Hypothetical protein	2
XP_001147565.1 (235)	*Pan troglodytes* (E)	Similar to RP11-346E17.3	2
XP_001253313.1 (176)	*Bos taurus* (E)	Similar to RP11-346E17.3	1
Q5W0N0 (161)	*Homo sapiens* (E)	Uncharacterized protein	2
AAI30405.1 (127)	*Homo sapiens* (E)	C9orf57 protein	2

Table 6.2b. List of proteins containing the 31 amino acid residue FYE repeat.

GENE_ID (number of residues)	Organism	Description	Number of FYE repeats
NP_083939.1 (660)	*Mus musculus* (E)	Eukaryotic translation elongation factor 1 delta	1
NP_001013122.1 (650)	*Rattus norvegicus* (E)	Eukaryotic translation elongation factor 1 delta	1
XP_856754.1 (611)	*Canis lupus familiaris* (E)	Eukaryotic translation elongation factor 1 delta	1
BAE01260.1 (669)	*Macaca fascicularis* (E)	Unnamed protein product	1
AAH00678.2 (550)	*Homo sapiens* (E)	Eukaryotic translation elongation factor 1 delta	1
XP_532345.2 (634)	*Canis lupus familiaris* (E)	Eukaryotic translation elongation factor 1 delta	1
XP_519999.2 (622)	*Pan troglodytes* (E)	Eukaryotic translation elongation factor 1 delta	1
EAW82231.1 (623)	*Homo sapiens* (E)	Eukaryotic translation elongation factor 1 delta	1
AAQ15199.1 (632)	*Homo sapiens* (E)	Eukaryotic translation elongation factor 1 delta	1
NP_115754.2 (647)	*Homo sapiens* (E)	Eukaryotic translation elongation factor 1 delta	1
BAB14925.1 (647)	*Homo sapiens* (E)	Unnamed protein product	1
AAH07847.1 (647)	*Homo sapiens* (E)	Eukaryotic translation elongation factor 1 delta	1
AAP36729.1 (648)	*Homo sapiens* (E)	Eukaryotic translation elongation factor 1 delta	1
XP_594628.3 (637)	*Bos taurus* (E)	Eukaryotic translation elongation factor 1 delta	1
XP_001232628.1 (679)	*Gallus gallus* (E)	Eukaryotic translation elongation factor 1 delta	1
CAI21007 (554)	*Danio rerio* (E)	Eukaryotic translation elongation factor 1 delta	2

Table 6.2c. List of proteins containing the 34 amino acid residue VHMM repeat.

GENE_ID (number of residues)	Organism	Description	Number of VHMM repeats
NP_001072997.2 (387)	*Homo sapiens* (E)	Hypothetical protein	1 + 2 (tandem)
XP_001139388.1 (250)	*Pan troglodytes* (E)	Similar to NY-REN-7 protein, partial	2
XP_001134706.1 (299)	*Pan troglodytes* (E)	Hypothetical protein	1 + 2 (tandem)
AAH28606.1 (261)	*Homo sapiens* (E)	LOC202134 protein	1 + 2 (tandem)
AAI01340.1 (144)	*Homo sapiens* (E)	LOC653316 protein	1
BAA34472.1 (334)	*Homo sapiens* (E)	KIAA0752 protein	1 + 2 (tandem)
AAD42863.1 (310)	*Homo sapiens* (E)	AF155097_1 NY-REN-7 antigen	1 + 2 (tandem)
NP_775934.3 (310)	*Homo sapiens* (E)	Hypothetical protein	1 + 2 (tandem)
BAF82207.1 (310)	*Homo sapiens* (E)	Unnamed protein product	3
NP_001072995.1 (114)	*Homo sapiens* (E)	Hypothetical protein	1

Table 6.2d. List of proteins containing the 34 amino acid residue TQG repeat.

GENE_ID (number of residues)	Organism	Description	Number of TQG repeats
Q8WWL7 (1395)	*Homo sapiens* (E)	G2/mitotic-specific cyclin-B3	2
NP_149020.2 (1395)	*Homo sapiens* (E)	Cyclin B3 isoform 3	2
CAC40024.1 (1395)	*Homo sapiens* (E)	Cyclin B3	2
XP_521063.2 (899)	*Pan troglodytes* (E)	Cyclin B3, partial	2
EAW89921.1 (1257)	*Homo sapiens* (E)	Cyclin B3, isoform CRA_b	2
NP_001005763.1 (1330)	*Canis lupus familiaris* (E)	Cyclin B3	1
BAF85143 (1395)	*Homo sapiens* (E)	Unnamed protein product	2

Table 6.2e. List of proteins containing the 51 amino acid residue PES repeat.

GENE_ID (number of residues)	Organism	Description	Number of PES repeats
Q8WWL7 (1395)	*Homo sapiens* (E)	G2/mitotic-specific cyclin-B3	2
NP_149020.2 (1395)	*Homo sapiens* (E)	Cyclin B3 isoform 3	2
CAC40024.1 (1395)	*Homo sapiens* (E)	Cyclin B3	2
XP_521063.2 (899)	*Pan troglodytes* (E)	Cyclin B3, partial	1
EAW89921.1 (1257)	*Homo sapiens* (E)	Cyclin B3, isoform CRA_b	2
NP_001005763.1 (1330)	*Canis lupus familiaris* (E)	Cyclin B3	1
BAF85143 (1395)	*Homo sapiens* (E)	Unnamed protein product	2

Table 6.2f. List of proteins containing the 34 amino acid residue HTQ repeat.

GENE_ID (number of residues)	Organism	Description	Number of HTQ repeats
XP_374705.3 (683)	*Homo sapiens* (E)	Similar to PKD 1 like 3	1+3 (tandem)
NP_001078865 (683)	*Homo sapiens* (E)	Similar to PKD 1 like 3	1+3 (tandem)
EAL23925 (687)	*Homo sapiens* (E)	Similar to PKD 1 like 3	1+3 (tandem)

Table 6.2g. List of proteins containing the 34 amino acid residue PTT repeat.

GENE_ID (number of residues)	Organism	Description	Number of PTT repeats
XP_374705.3 (683)	*Homo sapiens* (E)	Similar to PKD 1 like 3	3 (tandem)+3
EAL23926.1 (437)	*Homo sapiens* (E)	Similar to PKD 1 like 3	3 (tandem)+3
NP_001078865 (683)	*Homo sapiens* (E)	Similar to PKD 1 like 3	3 (tandem)
EAL23925 (687)	*Homo sapiens* (E)	Similar to PKD 1 like 3	3 (tandem)+3

Table 6.2h. List of proteins containing the 34 amino acid residue FSQ repeat.

GENE_ID (number of residues)	Organism	Description	Number of FSQ repeats
NP_008917.3 (778)	*Homo sapiens* (E)	Melanoma antigen family D 1	3 (tandem)
BAD51991.1 (562)	*Macaca fascicularis* (E)	Melanoma antigen, family D1	3 (tandem)
NP_001005333.1 (834)	*Homo sapiens* (E)	Melanoma antigen family D 1	3 (tandem)
AAG09704.1 (778)	*Homo sapiens* (E)	Melanoma antigen-encoding gene	3 (tandem)
EAW62896.1 (778)	*Homo sapiens* (E)	Melanoma antigen family D 1	3 (tandem)
NP_001001860.1 (778)	*Sus scrofa* (E)	Melanoma antigen family D 1	3 (tandem)
BAB84918.1 (521)	*Homo sapiens* (E)	FLJ00163 protein	3 (tandem)
BAE22540.1 (775)	*Mus musculus* (E)	Unnamed protein product	3 (tandem)
BAD90336.1 (798)	*Mus musculus* (E)	Melanoma antigen-encoding gene	3 (tandem)
XP_538044.2 (555)	*Canis lupus familiaris* (E)	Similar to melanoma antigen family D 1	3 (tandem)
AAH31461.1 (775)	*Mus musculus* (E)	Melanoma antigen, family D 1	3 (tandem)
NP_062765.1 (775)	*Mus musculus* (E)	Melanoma antigen family D 1	3 (tandem)
BAE27491.1 (775)	*Mus musculus* (E)	Unnamed protein product	3 (tandem)
AAK01203.1 (769)	*Mus musculus* (E)	Melanoma antigen-encoding gene	3 (tandem)
Q9ES73 (775)	*Rattus norvegicus* (E)	Melanoma-associated antigen D 1	3 (tandem)
NP_445861.1 (775)	*Rattus norvegicus* (E)	Melanoma antigen family D 1	3 (tandem)
NP_001039590.1 (353)	*Bos taurus* (E)	Hypothetical protein	3 (tandem)
AAH16438.1 (550)	*Mus musculus* (E)	Melanoma antigen-encoding gene	3 (tandem)

Table 6.2i. List of proteins containing the 36 amino acid residue PEG repeat.

GENE_ID (number of residues)	Organism	Description	Number of PEG repeats
NP_005453.2 (1142)	*Homo sapiens* (E)	Melanoma antigen family C1	3
XP_942006.1 (243)	*Homo sapiens* (E)	Similar to Melanoma-associated antigen C1	2
AAC24227.1 (1142)	*Homo sapiens* (E)	Cancer/testis antigen CT7	3
AAC18837.1 (1142)	*Homo sapiens* (E)	Melanoma-associated antigen	3
CAI42087.1 (118)	*Homo sapiens* (E)	Melanoma antigen family C1	1
XP_001126506.1 (801)	*Homo sapiens* (E)	Similar to Melanoma-associated antigen C1	2

Table 6.2j. List of proteins containing the 42 amino acid residue SSC repeat.

GENE_ID (number of residues)	Organism	Description	Number of SSC repeats
XP_001127353.1 (299)	*Homo sapiens* (E)	Hypothetical protein	1
EAW61038.1 (420)	*Homo sapiens* (E)	Isoform CRA_c	3
AAI12925 (261)	*Homo sapiens* (E)	Unnamed protein	2
XP_001104785.1 (159)	*Macaca mulatto* (E)	Hypothetical protein	1
EAW69824.1 (351)	*Homo sapiens* (E)	Isoform CRA_c	2
XP_934335 (366)	*Homo sapiens* (E)	Hypothetical protein	3
EAW69822.1 (396)	*Homo sapiens* (E)	Isoform CRA_a	3
EAW83662.1 (380)	*Homo sapiens* (E)	Isoform CRA_c	1
EAW69823.1 (455)	*Homo sapiens* (E)	Isoform CRA_b	3
XP_934335.2 (366)	*Homo sapiens* (E)	Hypothetical protein	2
EAW83660.1 (450)	*Homo sapiens* (E)	Isoform CRA_a	1
XP_530161.2 (225)	*Pan troglodytes* (E)	Hypothetical protein	1
NP_001004321.2 (259)	*Homo sapiens* (E)	Hypothetical protein	1
BAC87609.1 (259)	*Homo sapiens* (E)	Unnamed protein product	1

Table 6.2k. List of proteins containing the YCL 42 amino acid residue repeat.

GENE_ID (number of residues)	Organism	Description	Number of YCL repeats
NP_060880.3 (748)	*Homo sapiens* (E)	Holliday junction recognition protein	2
XP_516170.2 (744)	*Pan troglodytes* (E)	Hypothetical protein isoform 2	2
XP_001151551.1 (745)	*Pan troglodytes* (E)	Hypothetical protein isoform 1	2
EAW71064.1 (748)	*Homo sapiens* (E)	Hypothetical protein, isoform CRA_a	2
BAC11221.1 (748)	*Homo sapiens* (E)	Unnamed protein product	2
EAW71065.1 (524)	*Homo sapiens* (E)	Hypothetical protein, isoform CRA_b	2
XP_001065693.1 (672)	*Rattus norvegicus* (E)	Hypothetical protein	1
XP_237403.4 (672)	*Rattus norvegicus* (E)	Hypothetical protein	1
XP_001110520.1 (858)	*Macaca mulatta* (E)	Hypothetical protein	2
BAD36742.1 (650)	*Mus musculus* (E)	Fetal liver expressing gene 1	2
NP_941054.1 (667)	*Mus musculus* (E)	Hypothetical protein	2
XP_874813.2 (811)	*Bos taurus* (E)	Hypothetical protein	2 (tandem)
BAE02517.1 (465)	*Macaca fascicularis* (E)	Unnamed protein product	2
NP_766093.1 (591)	*Mus musculus* (E)	Hypothetical protein	1
XP_001005388.1 (561)	*Mus musculus* (E)	Hypothetical protein	1
BAC27950.1 (388)	*Mus musculus* (E)	Unnamed protein product	1
AAH62125.1 (401)	*Mus musculus* (E)	Hypothetical protein	1

Table 6.2l. List of proteins containing the 43 amino acid residue VSR repeat.

GENE_ID (number of residues)	Organism	Description	Number of VSR repeats
XP_374142.3 (1015)	*Homo sapiens* (E)	Hypothetical protein	1
XP_001083163.1 (1011)	*Macaca mulatta* (E)	Hypothetical protein	1
XP_001069975.1 (1017)	*Rattus norvegicus* (E)	Similar to proteoglycan 4	1
XP_896829.1 (1071)	*Mus musculus* (E)	Hypothetical protein	1
BAE26839.1 (1006)	*Mus musculus* (E)	Unnamed protein product	1
XP_001502524 (1019)	*Equus caballus* (E)	Hypothetical protein	1

Table 6.2m. List of proteins containing the 54 amino acid residue ALPG repeat.

GENE_ID (number of residues)	Organism	Description	Number of ALPG repeats
XP_896829.1 (1071)	*Mus musculus* (E)	Hypothetical protein	1
BAE26839.1 (1006)	*Mus musculus* (E)	Unnamed protein product	1
XP_001069975.1 (1017)	*Rattus norvegicus* (E)	Similar to proteoglycan 4	1
XP_001083163.1 (1011)	*Macaca mulatta* (E)	Hypothetical protein	1
XP_374142.3 (1015)	*Homo sapiens* (E)	Hypothetical protein	1
XP_001502524 (1019)	*Equus caballus* (E)	Hypothetical protein	1

Table 6.2n. List of proteins containing the 43 amino acid residue SVT repeat.

GENE_ID (number of residues)	Organism	Description	Number of SVT repeats
XP_499019.3 (376)	*Homo sapiens* (E)	Hypothetical protein	4 (tandem) + 2

Table 6.2o. List of proteins containing the 49 amino acid residue CDxD repeat.

GENE_ID (number of residues)	Organism	Description	Number of CDxD repeats
NP_003226.4 (2768)	*Homo sapiens* (E)	Thyroglobulin precursor	2
AAD50912.2 (1124)	*Homo sapiens* (E)	Thyroglobulin precursor	2
BAD92396.1 (1574)	*Homo sapiens* (E)	Thyroglobulin precursor	2
AAB53204.1 (2768)	*Mus musculus* (E)	Thyroglobulin precursor	2
O08710 (2766)	*Mus musculus* (E)	Thyroglobulin precursor	2
EAW92157.1 (2768)	*Homo sapiens* (E)	Thyroglobulin, isoform	2
CAA29104.1 (2767)	*Homo sapiens* (E)	Thyroglobulin precursor	2
AAC51924.1 (2768)	*Homo sapiens* (E)	Thyroglobulin precursor	2
P01266 (2768)	*Homo sapiens* (E)	Thyroglobulin precursor	2
AAC32269.1 (2766)	*Mus musculus* (E)	Thyroglobulin precursor	2
AAC32268.1 (2766)	*Mus musculus* (E)	Thyroglobulin precursor	2
NP_033401.2 (2766)	*Mus musculus* (E)	Thyroglobulin precursor	2
NP_112250.1 (2768)	*Rattus norvegicus* (E)	Thyroglobulin precursor	2
NP_776308.1 (2769)	*Bos taurus* (E)	Thyroglobulin precursor	2
NP_001041569.1 (2762)	*Canis lupus familiaris* (E)	Thyroglobulin precursor	1
AAF34909.1 (2768)	*Rattus norvegicus* (E)	Thyroglobulin precursor	2
CAF89701.1 (2122)	*Tetraodon nigroviridis* (E)	Unnamed protein product	1
CAA26183.1 (967)	*Rattus norvegicus* (E)	Unnamed protein product	1

Table 6.2p. List of proteins containing the 50 amino acid residue GGF repeat.

GENE_ID (number of residues)	Organism	Description	Number of GGF repeats
XP_941683.2 (4516)	*Homo sapiens* (E)	Similar to mucin 19	3 (tandem)
XP_497341.3 (7328)	*Homo sapiens* (E)	Similar to mucin 19	3 (tandem)

Table 6.2q. List of proteins containing the 52 amino acid residue NYS repeat.

GENE_ID (number of esidues)	Organism	Description	Number of NYS repeats
NP_048536.2 (1299)	*Homo sapiens* (E)	Similar to SWI/SNF chromatin remodeling complex subunit OSA2	4 (tandem)

Table 6.2r. List of proteins containing the 53 amino acid residue RPE repeat.

GENE_ID (number of residues)	Organism	Description	Number of RPE repeats
NP_835260.2 (2839)	*Homo sapiens* (E)	PDZ domain containing 2	1
AAK07661.1 (2641)	*Homo sapiens* (E)	PDZ domain containing protein AIPC	1
BAA20760.2 (2847)	*Homo sapiens* (E)	PDZ signaling protein	1
O15018 (2839)	*Homo sapiens* (E)	PDZ domain containing protein 3	1
EAX10777.1 (2665)	*Homo sapiens* (E)	Isoform CRA_c	1
XP_526957.2 (2443)	*Pan troglodytes* (E)	PDZ domain containing 2	1
XP_536512.2 (2601)	*Canis lupus familiaris* (E)	Similar to PDZ domain containing 3 isoform a	1
XP_871254.2 (2803)	*Bos taurus* (E)	Similar to KIAA0300	1

The proteins are represented by their corresponding GENE_ID along with the number of amino acid residues indicated in brackets in the first column. The organism and corresponding phylogeny are indicated in the second column; 'E' represents Eukaryota. The third column contains the description of the proteins containing the repeats identified in the present work. The fourth column represents exclusively the total number of novel repeats identified in this work.

Figure 6.1a: Multiple sequence alignment of 58 amino acid residue GPA domain.

```
Secondary structure                LLLLLL   LLL    LLLLL
XP_001174244.1(105-155)    APLGPAWASRRPLQAQIVLKSASPGPAPASRRPLQAQVVVKSAWN-W
XP_520604.2(92-148)        APLGPAWASRRPLQAQIVLKSASPGPAPASRRPLQAQVVVKSAWN-W
EAL23895.1(26-83)          GCPGPALASRRPLQAQVVLKSASPGPAPASRQPLWVQNFLESASPGP
NP_001013707.1(7-64)       ASPGPAWASRRPLQAQVVLKSASPGPAPASQQASSFGSAPAQLPPAF
                           .  *** *********:**************::.        .
consensus/80%              us.GPAhASRRPLQAQlVLKSASPGPAPASpps..htsh.tph.s..


Secondary structure
XP_001174244.1(105-155)    AWKSS------ 51
XP_520604.2(92-148)        AWKSSKSAFPG 57
EAL23895.1(26-83)          APPASQWPLSA 58
NP_001013707.1(7-64)       VDPELSPAMLL 58
                           .
consensus/80%              s..t.......
```

Figure 6.1b: Multiple sequence alignment of 61 amino acid residue RxH domain.

```
Secondary structure              CCCCCCCCCCCCCCC CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC
XP_892828.1_1(1811-1871)         KSLPSPQAAHKMLS-KAVSHRLHIADQEEPKNTAGDTSKPPQCVPE
AAI15887.1_1(357-417)            KSLPSPQAAHKMLS-KAVSHRLHIADQEEPKNTAGDTSKPPQCVPE
XP_912272.1_1(1864-1924)         KSLPSPQAAHKMLS-KAVSHRLHIADQEEPKNTAGDTSKPPQCVPE
BAC65522.1_1(419-479)            KSLPSPQAAHKMLS-KAVSHRLHIADQEEPKNTAGDTSKPPQCVPE
XP_981908.1_1(1868-1928)         KSLPSPQAAHKMLS-KAVSHRLHIADQEEPKNTAGDTSKPPQCVPE
NP_001074533.1_1(1863-1923)      KSLPSPQAAHKMLS-KAVSHRLHIADQEEPKNTAGDTSKPPQCVPE
NP_075229.1_1(1829-1889)         KTLPSPQASHKMFS-KAVSHRLHIADQEEPKNTAGDTPKPPQCVPE
XP_871254.2_1(1874-1933)         KALVSPQASHKMLS-KVASHRFHTADHEELDKTAAAAPQSPQCV-E
BAA20760.2_1(1923-1983)          KPLISPQTSHKTLS-KAVSQRLHVADHEDPDRNTTAAPRSPQCVLE
XP_526957.2_1(1518-1578)         KPLISPQTSHKTLS-KAVSQRLHVADHEDPDRNTTAAPRSPQCVLE
NP_835260.2_1(1915-1975)         KPLISPQTSHKTLS-KAVSQRLHVADHEDPDRNTTAAPRSPQCVLE
O15018_1(1915-1975)              KPLISPQTSHKTLS-KAVSQRLHVADHEDPDRNTTAAPRSPQCVLE
AAK07661.1_1(1717-1777)          KPLISPQTSHKTLS-KAVSQRLHVADHEDPDRNTTAAPRSPQCVLE
EAX10777.1_1(1741-1801)          KPLISPQTSHKTLS-KAVSQRLHVADHEDPDRNTTAAPRSPQCVLE
XP_536512.2_1(1656-1715)         KPLISPQASHRMLS-KAVAHRVHAPEHLEPGQDGPASPRPPPSGPE
NP_001074533.1_2(1928-1989)      RTITSPLTSPKPLPEQGANNRFHMAVYLESDTSCPATSRPPRYGPE
XP_912272.1_2(1929-1990)         RTITSPLTSPKPLPEQGANNRFHMAVYLESDTSCPATSRPPRYGPE
AAI15887.1_2(422-483)            RTITSPLTSPKPLPEQGANNRFHMAVYLESDTSCPATSRPPRYGPE
XP_981908.1_2(1933-1994)         RTITSPLTSPKPLPEQGANNRFHMAVYLESDTSCPATSRPPRYGPE
XP_892828.1_2(1876-1937)         RTITSPLTSPKPLPEQGANNRFHMAVYLESDTSCPATSRPPRYGPE
BAC65522.1_2(484-545)            RTITSPLTSPKPLPEQGANNRFHMAVYLESDTSCPATSRPPRYGPE
NP_075229.1_2(1894-1955)         RTFTSPLTSPKLLPEQGANSRFHMAVYLESDTSCPTTSRSPRSGPE
BAA20760.2_2(1988-2049)          RTFVSPLTSPKPVPEQGMWSRFHMAVLSEPDRGCPTTPKSPKCRAE
NP_835260.2_2(1980-2041)         RTFVSPLTSPKPVPEQGMWSRFHMAVLSEPDRGCPTTPKSPKCRAE
O15018_2(1980-2041)              RTFVSPLTSPKPVPEQGMWSRFHMAVLSEPDRGCPTTPKSPKCRAE
AAK07661.1_2(1782-1843)          RTFVSPLTSPKPVPEQGMWSRFHMAVLSEPDRGCPTTPKSPKCRAE
EAX10777.1_2(1806-1867)          RTFVSPLTSPKPVPEQGMWSRFHMAVLSEPDRGCPTTPKSPKCRAE
XP_526957.2_2(1583-1644)         RTFVSPLTSPKPLPEQGMWSRFHMAVLSEPDRGCPTTPKSPKCRAE
XP_871254.2_2(1938-1999)         RMFVSPVTTPKTLPEQGGCGRLHPAVHAEPDRGFPAAPSPPKCGPE
                                 : : ** :: : .. :    *.* .   :       :. .*    *
consensus/80%                    +sh.SP.su.K.ls.puh.pRhHhAs..-scpssssss+sPphh.E
```

```
Secondary structure             CCCCCCCCCCCCCCCC
XP_892828.1_1(1811-1871)        SKPPLAASGSLRTSAS 61
AAI15887.1_1(357-417)           SKPPLAASGSLRTSAS 61
XP_912272.1_1(1864-1924)        SKPPLAASGSLRTSAS 61
BAC65522.1_1(419-479)           SKPPLAASGSLRTSAS 61
XP_981908.1_1(1868-1928)        SKPPLAASGSLRTSAS 61
NP_001074533.1_1(1863-1923)     SKPPLAASGSLRTSAS 61
NP_075229.1_1(1829-1889)        SKPPQAALGSLRTSAS 61
XP_871254.2_1(1874-1933)        GKLPPGTPGSLKPSAS 60
BAA20760.2_1(1923-1983)         SKPPLATSGPLKPSVS 61
XP_526957.2_1(1518-1578)        SKPPLATSGPLKPSVS 61
NP_835260.2_1(1915-1975)        SKPPLATSGPLKPSVS 61
O15018_1(1915-1975)             SKPPLATSGPLKPSVS 61
AAK07661.1_1(1717-1777)         SKPPLATSGPLKPSVS 61
EAX10777.1_1(1741-1801)         SKPPLATSGPLKPSVS 61
XP_536512.2_1(1656-1715)        SGPP-ATPGSPKAPAE 60
NP_001074533.1_2(1928-1989)     GKVPHANSGSVSPSAS 62
XP_912272.1_2(1929-1990)        GKVPHANSGSVSPSAS 62
AAI15887.1_2(422-483)           GKVPHANSGSVSPSAS 62
XP_981908.1_2(1933-1994)        GKVPHANSGSVSPSAS 62
XP_892828.1_2(1876-1937)        GKVPHANSGSVSPSAS 62
BAC65522.1_2(484-545)           GKVPHANSGSVSPSAS 62
NP_075229.1_2(1894-1955)        GKAPHANSGSASPPAS 62
BAA20760.2_2(1988-2049)         GRAPRADSGPVSPAAN 62
NP_835260.2_2(1980-2041)        GRAPRADSGPVSPAAN 62
O15018_2(1980-2041)             GRAPRADSGPVSPAAN 62
AAK07661.1_2(1782-1843)         GRAPRADSGPVSPAAN 62
EAX10777.1_2(1806-1867)         GRAPRADSGPVSPAAN 62
XP_526957.2_2(1583-1644)        GRAPRADSGPVSPAAN 62
XP_871254.2_2(1938-1999)        SRAPLASPGPASPAAT 62
                                .  *  .  *.  ...
consensus/80%                   u+sPhAsSGslpsusS
```

Figure 6.1c: Multiple sequence alignment of 68 amino acid residue GLG domain.

```
Secondary structure
XP_496331.2_2(393-460)      PRAPAARPGHLLSCSAACWPQRTRVRRPPGLGISSPAAPPAGHREPA
XP_496331.2_4(325-392)      PRAPAARPGHLLSCSAACWPQRTRVRRPPGLGISSPAAPPAGHREPA
XP_496331.2_5(145-212)      PRAPAARPGHLLSCSTACWPQRTRVRRPPGLGISSPAAPPAGHREPA
XP_496331.2_3(213-280)      PRAPAAWPGHLLSCSAACWPQRTRVRRPPGLGISSPAAPPAGHREPA
XP_496331.2_6(505-572)      PRAPAAWPGHLLSCSAACWPQRTRVRRPPGLGISSPAAPPAGHREPA
XP_496331.2_1(77-144)       PPEQAQAPAPELSCSAACWPQRAHVRRPPGLGISPPAAPPAGHREPA
XP_496331.2_7(437-504)      EPACAGRQAWASPLLQRRLLATENPRVPPGLGISSPAAPPAGHREPA
                                 *   .   .          . * *******.************
consensus/80%               P.AsAuhPuahLSCSsACWPQRs+VRRPPGLGISSPAAPPAGHREPA


Secondary structure              HHHHHHHHHHHH
XP_496331.2_2(393-460)      CAGRQAWASPLLQRRLLATEN 68
XP_496331.2_4(325-392)      CAGRLAWASPLLQRRLLATEN 68
XP_496331.2_5(145-212)      CAGRQAWASPLLQHRLLATEN 68
XP_496331.2_3(213-280)      CAGRQAWASPLLQRRLLATEN 68
XP_496331.2_6(505-572)      CAGRQAWASP---RALVPPLT 65
XP_496331.2_1(77-144)       CAGRQAWASPLLQRRLLATEN 68
XP_496331.2_7(437-504)      CAGRQAWASPLLQRRLLATEN 68
                            **** *****   : *:.. .
consensus/80%               CAGRQAWASPLLQRRLLATEN
```

Figure 6.1d: Multiple sequence alignment of 71 amino acid residue SAS domain.

```
Secondary structure           HHHHHHHHHHHHHHHHHHHH  HHHHHHHHHHHHHHHHHHHHHHHHH
CAD38776.1(109-179)       ASSASKNKAQNITAPESEAICWQLLSRCFRSWRHLVKRQREPAAAAVALGRW
XP_943296.2(489-559)      ASSASKNKAQNITAPESEAICWQLLSRCFRSWRHLVKRQREPAAAAVALGRW
CAI15880.1(104-174)       ASSASKNKAQNITAPESEAICWQLLSRCFRSWRHLVKRQREPAAAAVALGRW
XP_209234.6(489-559)      ASSASKNKAQNITAPESEAICWQLLSRCFRSWRHLVKRQREPAAAAVALGRW
                          ***************************************************
consensus/80%             ASSASKNKAQNITAPESEAICWQLLSRCFRSWRHLVKRQREPAAAAVALGRW


Secondary structure       HH HHHHHHHHHHHHHH
CAD38776.1(109-179)       QLLQKCLQALWLREAQLEA 71
XP_943296.2(489-559)      QLLQKCLQALWLREAQLEA 71
CAI15880.1(104-174)       QLLRKCLQALWLREAQLEA 71
XP_209234.6(489-559)      QLLRKCLQALWLREAQLEA 71
                          ***:***************
consensus/80%             QLLpKCLQALWLREAQLEA
```

Figure 6.1e: Multiple sequence alignment of 73 amino acid residue WKRK domain.

```
Secondary structure                                          HHHH
XP_001110275.1_1(129-201)    APGVDPRPSCRSLCWKRKREWSDQSEE----EPEKAL----SPEPEE
XP_001110275.1_2(202-274)    ALGVDPSPSCRSVCWKRKREWSDQSEE----EPEKAL----SPEPEE
XP_001115801.1_1(46-118)     ALGVDPSPSCRSLCWKRKREWLDQSDE----EPEKEL----SPEPEE
XP_001142986.1_2(87-162)     APGVDPSPLHRSFCWKRKMEWWDESEE----SL-EELRKVLAPEPEE
XP_001156328.1_2(87-161)     APGVDPSPPHRSFCWKRKMEWWDESEE----S--EELRKVLAPEPEE
XP_001134912.1_2(127-202)    APGVDPSPPHRSFCWKRKMEWWDESEE----SE-EELRKVLAPEPEE
XP_935532_1(87-163)          APGVDPSPPHRSFCWKRKMEWWDESEE----SLEEEPRKVLAPEPEE
NP_001092905.1_2(87-163)     APGVDPSPPHRSFCWKRKMEWWDESEE----SLEEEPRKVLAPEPEE
EAW50237.1_1(30-106)         APGVDPSPPHRSFCWKRKMEWWDESEE----SLEEEPRKVLAPEPEE
XP_499356.1_1(87-163)        APGVDPSPPHRSFCWKRKMEWWDESEE----SLEEEPRKVLAPEPEE
NP_778234.2_1(87-162)        APGVDPSPPHRSFCWKRKMEWWDKSEE----SE-EEPRKVLAPEPEE
AAM62309.1_1(87-162)         APGVDPSPPHRSFCWKRKMEWWDKSEE----SE-EEPRKVLAPEPEE
CAB70665.1_1(59-134)         APGVDPSPPHRSFCWKRKMEWWDKSEE----SE-EEPRKVLAPEPEE
XP_001127201.1(87-163)       APGVDPSPPHRSFCWKRKREWWDESEE----SLEEEPRKVLAPEPEE
XP_499314_1(87-163)          APGVDPSPPHRSFCWKRKREWWDESEE----SLEEEPRKVLAPEPEE
XP_001130563.1_1(174-250)    APGVDPSPPHRSFCWKRKREWWDESEE----SLEEEPRKVLAPEPEE
EAX07879.1_1(87-163)         APGVDPSPPHRSFCWKRKREWWDESEE----SLEEEPRKVLAPEPEE
XP_001130493.1_1(143-219)    APGVDPSPPHRSFCWKRKREWWDESEE----SLEEEPRKVLAPEPEE
XP_001152570.1_1(44-116)     APGVDPSPPRRSLGWKRKRECLDESDD----EPEQEL----APEPEE
XP_001152377.1_1(44-116)     APGVDPSPPRRSLGWKRKRECLDESDD----EPEQEL----APEPEE
XP_001150008.1_1(44-116)     APGVDPSPPRRSLGWKRKRECLDESDD----EPEKEL----APEPEE
XP_001149803.1_1(44-116)     APGVDPSPPRRSLGWKRKRECLDESDD----EPEKEL----APEPEE
XP_940755.1_1(44-116)        APGVDPSPPRRSLGWKRKRECLDESDD----EPEKEL----APEPEE
XP_496899.1_1(44-116)        APGVDPSPPRRSLGWKRKRECLDESDD----EPEKEL----APEPEE
AAH45636.1_1(36-108)         APGVDPSPPRRSLGWKRKRECLDESDD----EPEKEL----APEPEE
AAH56606_1(36-108)           APGVDPSPPRRSLGCKRKRECLDESDD----EPEKEL----APEPEE
AAH56606_2(109-181)          APGVDPSPPRRSLGCKRKRECLDESDD----EPEKEL----APEPEE
AAH56606_3(182-254)          APGVDPSPPRRSLGCKRKRECLDESDD----EPEKEL----APEPEE
AAH56606_4(255-327)          APGVDPSPPRRSLGCKRKRECLDESDD----EPEKEL----APEPEE
AAH56606_5(328-400)          APGVDPSPPRRSLGCKRKRECLDESDD----EPEKEL----APEPEE
XP_499348.2_1(36-108)        APGVDPSPPRRSLGCKRKRECLDESDD----EPEKEL----APEPEE
XP_371014.3(37-113)          APWIDPSPQPQSLGLKRKSEWSDESEE----ELEEELELERAPEPED
XP_001156328.1_1(14-86)      APGVDPSPPCRSLGWKRKREWSDESAE----QPEKEL----APEPEE
XP_001152762.1_1(54-126)     APGVDPSPPCRSLGWKRKKEWSDESAE----EPEKEL----APEPEE
XP_001156047.1_1(14-86)      APGVDPSPPCRSLGWKRKREWSDESAE----EPEKEL----APEPEE
XP_001135002.1_1(14-86)      APGVDPSPPCRSLGWKRKREWSDESAE----EPEKEL----APEPEE
XP_001142986.1_1(14-86)      APGVDPSPPCRSLGWKRKREWSDESAE----EPEKEL----APEPEE
NP_001092905.1_1(14-86)      APGVDPSPPCRSLGWKRKREWSDESAE----EPEKEL----APEPEE
XP_001130563.1_2(101-173)    APGVDPSPPCRSLGWKRKKEWSDESEE----EPEKEL----APEPEE
XP_001130493.1_2(70-142)     APGVDPSPPCRSLGWKRKKEWSDESEE----EPEKEL----APEPEE
EAX07879.1_2(14-86)          APGVDPSPPCRSLGWKRKKEWSDESEE----EPEKEL----APEPEE
XP_001127201.2(14-86)        APGVDPSPPCRSLGWKRKKEWSDESEE----EPEKEL----APEPEE
XP_499314_2(14-86)           APGVDPSPPCRSLGWKRKKEWSDESEE----EPEKEL----APEPEE
AAM62309.1_2(14-86)          APGVDPSPPCRSLGWKRKREWSDESEE----EPEKEL----APEPEE
NP_778234.2_2(14-86)         APGVDPSPPCRSLGWKRKREWSDESEE----EPEKEL----APEPEE
XP_935532_2(14-86)           APGVDPSPPCRSLGWKRKREWSDESEE----EPEKEL----APEPEE
XP_499356.1_2(14-86)         APGVDPSPPCRSLGWKRKREWSDESEE----EPEKEL----APEPEE
EAW94419.1_1(116-192)        APGVDPSPLCGSLCWKRKREWSDESEESSEEEPEKEL----APEPEE
EAW94419.1_2(32-115)         APGVDPSLPCRSLCWKRKREWSDQSEESSEEAPEKEL----APEPEE
                             *  :**     *.  *** *  *:* :       :      :****:
consensus/80%                APGVDPSPPpRShsWKRKhEh.DES--....p.EcE.....APEPEE
```

```
Secondary structure              EEEE     HHHHH          HHHHHHHHHH
XP_001110275.1_1(129-201)      TWVVETLCGLKMKLKQRR--VSPVLPEHHEAFNSQL 73
XP_001110275.1_2(202-274)      TWVVETLCGLKMKLKRRR--VSPVLPEHHEAFNRLL 73
XP_001115801.1_1(46-118)       TWVAETLCGLKMKLKRRR--VSLVLPEHHEAFNRLL 73
XP_001142986.1_2(87-162)       IWVAEMLCGLKMKLKRRR--VSLVLPEHHEAFNRLL 76
XP_001156328.1_2(87-161)       IWVAEMLCGLKMKLKRRR--VSLVLPEHHEAFNRLL 75
XP_001134912.1_2(127-202)      IWVAEMLCGLKMKLKRRR--VSLVLPEHHEAFNRLL 76
XP_935532_1(87-163)            IWVAEMLCGLKMKLKRRR--VSLVLPEHHEAFNRLL 77
NP_001092905.1_2(87-163)       IWVAEMLCGLKMKLKRRR--VSLVLPEHHEAFNRLL 77
EAW50237.1_1(30-106)           IWVAEMLCGLKMKLKRRR--VSLVLPEHHEAFNRLL 77
XP_499356.1_1(87-163)          IWVAEMLCGLKMKLKRRR--VSLVLPEHHEAFNRLL 77
NP_778234.2_1(87-162)          IWVAEMLCGLKMKLKRRR--VSLVLPEHHEAFNRLL 76
AAM62309.1_1(87-162)           IWVAEMLCGLKMKLKRRR--VSLVLPEHHEAFNRLL 76
CAB70665.1_1(59-134)           IWVAEMLCGLKMKLKRRR--VSLVLPEHHEAFNRLL 76
XP_001127201.1(87-163)         IWVAEMLCGLKMKLKRRR--VSLVLPEHHEAFNRLL 77
XP_499314_1(87-163)            IWVAEMLCGLKMKLKRRR--VLLVLPEHHEAFNRLL 77
XP_001130563.1_1(174-250)      IWVAEMLCGLKMKLKRRR--VSLVLPEHHEAFNRLL 77
EAX07879.1_1(87-163)           IWVAEMLCGLKMKLKRRR--VSLVLPEHHEAFNRLL 77
XP_001130493.1_1(143-219)      IWVVEMLCGLKMKLKRRR--VSLVLPEHHEAFNRLL 77
XP_001152570.1_1(44-116)       TWVAETLCGLKMKAKRRR--VSLVLPEYYEAFNRLL 73
XP_001152377.1_1(44-116)       TWVAETLCGLKMKAKRRR--VSLVLPEYYEAFNRLL 73
XP_001150008.1_1(44-116)       TWVAETLCGLKMKAKRRR--VSLVLPEYYEAFNRLL 73
XP_001149803.1_1(44-116)       TWVAETLCGLKMKAKRRR--VSLVLPEYYEAFNRLL 73
XP_940755.1_1(44-116)          TWVAETLCGLKMKAKRRR--VSLVLPEYYEAFNRLL 73
XP_496899.1_1(44-116)          TWVAETLCGLKMKAKRRR--VSLVLPEYYEAFNRLL 73
AAH45636.1_1(36-108)           TWVAETLCGLKMKAKRRR--VSLVLPEYYEAFNRLL 73
AAH56606_1(36-108)             TWVAETLCGLKMKAKRRR--VSLVLPEYYEAFNRLL 73
AAH56606_2(109-181)            TWVAETLCGLKMKAKRRR--VSLVLPEYYEAFNRLL 73
AAH56606_3(182-254)            TWVAETLCGLKMKAKRRR--VSLVLPEYYEAFNRLL 73
AAH56606_4(255-327)            TWVAETLCGLKMKAKRRR--VSLVLPEYYEAFNRLL 73
AAH56606_5(328-400)            TWVAETLCGLKMKAKRRR--VSLVLPEYYEAFNRLL 73
XP_499348.2_1(36-108)          TWVAETLCGLKMKAKRRR--VSLVLPEYYEAFNRLL 73
XP_371014.3(37-113)            TWVVETLCGLKMKLKRKR--ASSVLPEHHEAFNRLL 77
XP_001156328.1_1(14-86)        TWVVETMCGLTMKLKQQQ--VSSFLPEHHKDFNSQL 73
XP_001152762.1_1(54-126)       TWVVETMCGLTMKLKQQQ--VSPFLPEHHKDFNSQL 73
XP_001156047.1_1(14-86)        TWVVETMCGLTMKLKQQR--VSPILPEHHKDFNSQL 73
XP_001135002.1_1(14-86)        TWVVEMPCGLTMKLKQQR--VSPILPEHHKDFNSQL 73
XP_001142986.1_1(14-86)        TWVLETLCGLKMKLKQQR--VSPILPEHHKDFNSQL 73
NP_001092905.1_1(14-86)        TWVVEMLCGLKMKLKQQR--VSPILPEHHKGFNSQL 73
XP_001130563.1_2(101-173)      TWVVEMLCGLKMKLKQQR--VSPILPEHHKDFNSQL 73
XP_001130493.1_2(70-142)       TWVVEMLCGLKMKLKQQR--VSPILPEHHKDFNSQL 73
EAX07879.1_2(14-86)            TWVVETLCGLKMKLKQQR--VSPILPEHHKDFNSQL 73
XP_001127201.2(14-86)          TWVVETLCGLKMKLKQQR--VSPILPEHHKDFNSQL 73
XP_499314_2(14-86)             TWVVETLCGLKMKLKQQR--VSPILPEHHKDFNSQL 73
AAM62309.1_2(14-86)            TWVVEMLCGLKMKLKQQR--VSPILPEHHKDFNSQL 73
NP_778234.2_2(14-86)           TWVVETLCGLKMKLKQQR--VSPILLEHHKDFNSQL 73
XP_935532_2(14-86)             TWVVEMLCGLKMKLKQQR--VSSILPEHHKDFNSQL 73
XP_499356.1_2(14-86)           TWVVEMLCGLKMKLKQQR--VSSILPEHHKDFNSQL 73
EAW94419.1_1(116-192)          TWVAETLCGLKMKLKQWR--VSPVLPEHHETFNRLL 77
EAW94419.1_2(32-115)           TWVAEMLCGLKMKLKLKQRLVSFVLPEHHEDFNRLL 79
                                ** *  ***.** *  :  .  .* *::: **  *
consensus/80%                  hWVsEhLCGLKMKhKppR..VS.lLPEaacsFNp.L
```

Figure 6.1f: Multiple sequence alignment of 85 amino acid residue FSS domain.

```
Secondary structure
Q8WY91_2(277-362)              DKGLAQSPPSSSLTATPQKPSQSPSAPPADVTPKPATEAV--QSEHSD
XP_001128859.1_2(384-469)      DKGLAQSPPSSSLTATPQKPSQSPSAPPADVTPKPATEAV--QSEHSD
NP_057047.3_2(277-362)         DKGLAQSPPSSSLTATPQKPSQSPSAPPADVTPKPATEAV--QSEHSD
EAW71276.1_2(277-362)          DKGLAQSPPSSSLTATPQKPSQSPSAPPADVTPKPATEAV--QSEHSD
XP_516210.2_2(386-471)         DKGLAQSPPSSSLTATPQKPSQSPSAPPADVTPKPATEAV--QSEHSD
XP_001093438.1_2(142-227)      DKGLAQSPPGSSLTATPQKPSQSPSAPPADVTPKPATEAV--QSQHSD
AAH71896.1_1(31-115)           DKGLAQSPPSSSLTATPQKPSQSPSAPPADVTPKPATEAV--QSEHSD
XP_543333.2_2(332-417)         DKGLAQSPPSSSLTATPQKPSQSPSAPPTDVTPKPAAEAV--QSEHSD
AAH57963.2_2(133-218)          GNSVAQSPPSSSLTATPQKASQSPSAPPTDVTPKPAAEAV--QSEHSD
NP_001005564.1_2(266-351)      GNSVAQSPPSSSLTATPQKASQSPSAPPTDVTPKPAAEAV--QSEHSD
NP_080196.3_2(266-351)         GNSVAQSPPSSSLTATPQKASQSPSAPPTDVTPKPAAEAV--QSEHSD
BAE24509.1_2(266-351)          GNSVAQSPPSSSLTATPQKASQSPSAPPTDVTPKPAAEAV--QSEHSD
AAH66042.1_2(266-351)          GNSVAQSPPSSSLTATPQKASQSPSAPPTDVTPKPAAEAV--QSEHSD
NP_001033758.1_2(284-369)      SPPRACPSASSSLTATPQKPAQGASAPPTDVTPKPAAEAV--QSEHSD
NP_080196.3_1(163-247)         GQSLEKTPGDDPAAPLARG-QEEAQASATEADHQKASSS--TDAEGAD
AAH57963.2_1(30-114)           GQSLEKTPGDDPAAPLARG-QEEAQASATEADHQKASSS--TDAEGAD
BAE24509.1_1(163-247)          GQSLEKTPGDDPAAPLARG-QEEAQASATEADHQKASSS--TDAEGAD
AAH66042.1_1(163-247)          GQSLEKTPGDDPAAPLARG-QEEAQASATEADHQKASSS--TDAEGAD
NP_001005564.1_1(163-247)      GQSLERTPGDQ-AAPLARG-QEEAQVSATEADHQKASSSA-ADAGGAD
XP_516210.2_1(277-361)         QQALERTPGDGLATMVAGS-QGKAEASATDAGDESATSSI--EGGVTD
XP_001128859.1_1(275-359)      QQALERTPGDGLATMVAGS-QGKAEASATDAGDESATSSI--EGGVTD
NP_057047.3_1(168-252)         QQALERTPGDGLATMVAGS-QGKAEASATDAGDESATSSI--EGGVTD
Q8WY91_1(168-252)              QQALERTPGDGLATMVAGS-QGKAEASATDAGDESATSSI--EGGVTD
EAW71276.1_1(168-252)          QQALERTPGDGLATMVAGS-QGKAEASATDAGDESATSSI--EGGVTD
XP_001093438.1_1(33-117)       QQALERTPGDGLATTVAGS-QGKAEVSATDAGDESTTPSI--EGGVAD
XP_543333.2_1(222-307)         RQALEGTPGDGPAP-AAGG-QGEADASAMDAGGDSAAAAAPPDGGLVD
NP_001033758.1_1(169-253)      RQPLEGRAEDGPASAATSCSQGEAGTGAEDAGEEGATPA---DRGLVD
                                       . .   .  .      . . . :.  . :: :   :    *
consensus/80%                  tpultpoPssu.hs.sstt.ptpspAsssDss.csAspuh..puthsD

Secondary structure                            HHHHHHHHHH          HH
Q8WY91_2(277-362)              ASPMSINEVILSASGACKLIDSLHSYCFSSRQNKSQVCCL 86
XP_001128859.1_2(384-469)      ASPMSINEVILSASGACKLIDSLHSYCFSSRQNKSQVCCL 86
NP_057047.3_2(277-362)         ASPMSINEVILSASGACKLIDSLHSYCFSSRQNKSQVCCL 86
EAW71276.1_2(277-362)          ASPMSINEVILSASGACKLIDSLHSYCFSSRQNKSQVCCL 86
XP_516210.2_2(386-471)         ASPMSINEVILSASGACKLIDSLHSYCFSSRQNKSQVCCL 86
XP_001093438.1_2(142-227)      ASPMSINEVILSASGACKLIDSLHSYCFSSRQNKSQVCCL 86
AAH71896.1_1(31-115)           ASPMSINEVILSASGACKLIDSLHSYCFSSRQNKSQVCCL 85
XP_543333.2_2(332-417)         ASPMSINEVILSASGACKLIDSLHSYCFSSRQNKSQVCCL 86
AAH57963.2_2(133-218)          ASPMSINEVILSASGACKLIDSLHSYCFSARQNKSQVCCL 86
NP_001005564.1_2(266-351)      ANPMSINEVILSASGACKLIDSLHSYCFSARQNKSQVCCL 86
NP_080196.3_2(266-351)         ASPMSINEVILSASGACKLIDSLHSYCFSARQNKSQVCCL 86
BAE24509.1_2(266-351)          ASPMSINEVILSASGACKLIDSLHSYCFSARQNKSQVCCL 86
AAH66042.1_2(266-351)          ASPMSINEVILSASGACKLIDSLHSYCFSARQNKSQVCCL 86
NP_001033758.1_2(284-369)      ASPMSINEVILSASGACKLIDSLHSYCFSSRQSKSQVCCL 86
NP_080196.3_1(163-247)         KSGISMDDFTPPGSGACKFIGSLHSYSFSSKHTRERPSVP 85
AAH57963.2_1(30-114)           KSGISMDDFTPPGSGACKFIGSLHSYSFSSKHTRERPSVP 85
BAE24509.1_1(163-247)          KSGISMDDFTPPGSGACKFIGSLHSYSFSSKHTRERPSVP 85
AAH66042.1_1(163-247)          KSGISMDDFTPPGSGACKFIGSLHSYSFSSKHTRERPSVP 85
NP_001005564.1_1(163-247)      KSGISMDDFTPPGSGACKFIGSLHSYSFSSKHTRERPSVP 85
XP_516210.2_1(277-361)         KSGISMDDFTPPGSGACKFIGSLHSYSFSSKHTRERPSVP 85
XP_001128859.1_1(275-359)      KSGISMDDFTPPGSGACKFIGSLHSYSFSSKHTRERPSVP 85
NP_057047.3_1(168-252)         KSGISMDDFTPPGSGACKFIGSLHSYSFSSKHTRERPSVP 85
Q8WY91_1(168-252)              KSGISMDDFTPPGSGACKFIGSLHSYSFSSKHTRERPSVP 85
EAW71276.1_1(168-252)          KSGISMDDFTPPGSGACKFIGSLHSYSFSSKHTRERPSVP 85
XP_001093438.1_1(33-117)       KSGISMDDFTPPGSGACKFIGSLHSYSFSSKHTRERPSVP 85
XP_543333.2_1(222-307)         KSGISKDDFTPPGSGACKFIGSLHSYSFSSKHARERASVP 86
NP_001033758.1_1(169-253)      RSGVSADDFTPPGSGACKFIGSLHSYSFSSKHARERPAVP 85
                                . :* ::.  ..*****:*.*****.**::: :.: .
consensus/80%                  tSshShs-hh.suSGACKhIsSLHSYsFSS+ps+ppsss.
```

Figure 6.1g: Multiple sequence alignment of 109 amino acid residue LLE domain.

```
Secondary structure                    HHHHHHHHHHHHHHHHHHHHHH        HHHHHHH
BAA32323.2_2(361-472)          GKTLSVLLLEHKEDLIQCVTQLRPIMESLETAKEEFLTGTEKRVILNCC
Q9NUA8_2(347-458)              GKTLSVLLLEHKEDLIQCVTQLRPIMESLETAKEEFLTGTEKRVILNCC
NP_055685.2_2(347-458)         GKTLSVLLLEHKEDLIQCVTQLRPIMESLETAKEEFLTGTEKRVILNCC
CAI22041.1_2(347-453)          GKTLSVLLLEHKEDLIQCVTQLRPIMESLETAKEEFLTGTEKRVILNCC
XP_001164879.1_2(347-458)      GKTLSVLLLEHKEDLIQCVTQLRPIMEFLETAKEEFLTGTEKRVILNCC
XP_001164989.1_2(300-411)      GKTLSVLLLEHKEDLIQCVTQLRPIMEFLETAKEEFLTGTEKRVILNCC
XP_001164955.1_2(347-458)      GKTLSVLLLEHKEDLIQCVTQLRPIMEFLETAKEEFLTGTEKRVILNCC
XP_001101017.1_2(300-411)      GKTLSVLLLEHKEDLIQCVTQLRPIMEFLETAKEEFLTGTEKRVILNCC
XP_001101280.1_2(300-411)      GKTLSVLLLEHKEDLIQCVTQLRPIMEFLETAKEEFLTGTEKRVILNCC
XP_001101193.1_2(347-458)      GKTLSVLLLEHKEDLIQCVTQLRPIMEFLETAKEEFLTGTEKRVILNCC
NP_937891.1_2(337-448)         GKTLSVLLLEHKEDLIQCVTQLRPIVEFLETAKEEFLPDSEKRVIQSCC
BAD90181.1_2(313-424)          GKTLSVLLLEHKEDLIQCVTQLRPIVEFLETAKEEFLPDSEKRVIQSCC
XP_919018.2_1(78-189)          GKTLSVLLLEHKEDLIQCVTQLRPIVEFLETAKEEFLPDSEKRVIQSCC
XP_342954.3_2(335-446)         GKTLSVLLLEHKEDLIQCVTQLRPIVEFLETAKEEFLPDSEKRVIQICC
XP_544510.2_2(348-459)         GKTLSALLLEHKEDLIQCVTQLRPILEFLETAEEEFLTGTEKRVILNCC
XP_614579.2_2(343-454)         GKTLSVLLLEHKEDLIRCVTQLRPILEFLETVEEEFLPGSEKRVILNCC
AAI14608.1_1(237-346)          KHYQLNFLLENEGVFSDALMVTQDVLKKLEMCSE--IKGPQKEVILNCC
XP_001165023.1_1(237-346)      KHYQLNFLLENEGIFSDALMVTQDVLKKLEMCSE--IKGPQKEVILNCC
XP_001253042.1_1(234-343)      KHYQLNFLLENEKIFSDALLITQDVLKRLEECSE--IKDVQKQVILNCC
XP_001164989.1_1(190-298)      KHYQLNFLLENEGIFSDALMVTQDVLKKLEMCSE--IKGPQKEMIVKCF
XP_001164955.1_1(237-345)      KHYQLNFLLENEGIFSDALMVTQDVLKKLEMCSE--IKGPQKEMIVKCF
XP_001164879.1_1(237-345)      KHYQLNFLLENEGIFSDALMVTQDVLKKLEMCSE--IKGPQKEMIVKCF
BAA32323.2_1(251-359)          KHYQLNFLLENEGVFSDALMVTQDVLKKLEMCSE--IKGPQKEMIVKCF
Q9NUA8_1(237-345)              KHYQLNFLLENEGVFSDALMVTQDVLKKLEMCSE--IKGPQKEMIVKCF
NP_055685.2_1(237-345)         KHYQLNFLLENEGVFSDALMVTQDVLKKLEMCSE--IKGPQKEMIVKCF
CAI22041.1_1(237-345)          KHYQLNFLLENEGVFSDALMVTQDVLKKLEMCSE--IKGPQKEMIVKCF
XP_001101017.1_1(190-298)      KRYQLNFLLENEGIFSDALMVTQDVLKKLEVCSE--IKGPQKETVVKCL
XP_001101280.1_1(190-298)      KRYQLNFLLENEGIFSDALMVTQDVLKKLEVCSE--IKGPQKETVVKCL
XP_001101193.1_1(237-345)      KRYQLNFLLENEGIFSDALMVTQDVLKKLEVCSE--IKGPQKETVVKCL
XP_544510.2_1(238-346)         KHHQLNFLLENESIFSDALLVTKDVLKRLEECSE--IKGAQEETIMGCL
XP_614579.2_1(234-341)         KHYQLNFLLENEKIFSDALLITQDVLKRLEECSE--IKDVQKQTIMGCL
NP_937891.1_1(237-345)         KQDQLTLLLEHESVFSDAALLTPDTLRRLAGCSE--IEGPQKETIIECL
BAD90181.1_1(213-321)          KQDQLTLLLEHESVFSDAALLTPDTLRRLAGCSE--IEGPQKETIIECL
XP_342954.3_1(237-343)         KQDQLTLLLEYESVFSSAALLTPDMLRRLEGCSE--IEGPQKETVMECL
                                :    :*** :  :  .       :. *   .*  : . ::. :  *
consensus/80%                  t+h..shLLEpct.h.pslh.hpslhchLEhspE..lpGspKchIhpCh
```

```
Secondary structure                HHHHHHHHHHHH       HHHHHHHHHHHHH
BAA32323.2_2(361-472)          EGRTPKETIENLLHRMTEEKTLTAEGLVKLLQAVKTTFPNL
Q9NUA8_2(347-458)              EGRTPKETIENLLHRMTEEKTLTAEGLVKLLQAVKTTFPNL
NP_055685.2_2(347-458)         EGRTPKETIENLLHRMTEEKTLTAEGLVKLLQAVKTTFPNL
CAI22041.1_2(347-453)          EGRTPKETIENLLHRMTEEKTLTAEGLVKLLQAVKTTFPNL
XP_001164879.1_2(347-458)      EGRTPKETIENLLHRMTEEKTLTAEGLVKLLQAVKATFPNL
XP_001164989.1_2(300-411)      EGRTPKETIENLLHRMTEEKTLTAEGLVKLLQAVKATFPNL
XP_001164955.1_2(347-458)      EGRTPKETIENLLHRMTEEKTLTAEGLVKLLQAVKATFPNL
XP_001101017.1_2(300-411)      EGRTPKETIENLLHRMTEEKTLTAESLVKLLQAVKMTFPNL
XP_001101280.1_2(300-411)      EGRTPKETIENLLHRMTEEKTLTAESLVKLLQAVKMTFPNL
XP_001101193.1_2(347-458)      EGRTPKETIENLLHRMTEEKTLTAESLVKLLQAVKMTFPNL
NP_937891.1_2(337-448)         EGRTPKEMIENLLHRVTEEKTLPAKSLVKLLQAVRTAFPNL
BAD90181.1_2(313-424)          EGRTPKEMIENLLHRVTEEKTLPAKSLVKLLQAVRTAFPNL
XP_919018.2_1(78-189)          ESRTPKETIENLLHRVTEEKTLPAKSLVKLLQAVRTAFPNL
XP_342954.3_2(335-446)         EGRTPKETVENLLHRVTEEKTLPAKSLVKLLQAVRTAFPNL
XP_544510.2_2(348-459)         EGGTPKETMEKLLHRMSEDKTLTAESLVKLLQAVKPMSPNL
XP_614579.2_2(343-454)         EGSTPKETIEKLLHRMSEEKTLSAESLVKLLQAVKMTFPDL
AAI14608.1_1(237-346)          EGRTPKETIENLLHRMTEEKTLTAEGLVKLLQAVKTTFPNL
XP_001165023.1_1(237-346)      EGRTPKETIENLLHRMTEEKTLTAEGLVKLLQAVKATFPNL
XP_001253042.1_1(234-343)      EGSTPKETIEKLLHRMSEEKTLSAESLVKLLQAVKMTFPDL
XP_001164989.1_1(190-298)      EGEGGRSAFQRILGKVREE-SLDVQTVVSLLRLYQYSNPAV
XP_001164955.1_1(237-345)      EGEGGRSAFQRILGKVREE-SLDVQTVVSLLRLYQYSNPAV
XP_001164879.1_1(237-345)      EGEGGRSAFQRILGKVREE-SLDVQTVVSLLRLYQYSNPAV
BAA32323.2_1(251-359)          EGEGGHSAFQRILGKVREE-SLDVQTVVSLLRLYQYSNPAV
Q9NUA8_1(237-345)              EGEGGHSAFQRILGKVREE-SLDVQTVVSLLRLYQYSNPAV
NP_055685.2_1(237-345)         EGEGGHSAFQRILGKVREE-SLDVQTVVSLLRLYQYSNPAV
CAI22041.1_1(237-345)          EGEGGHSAFQRILGKVREE-SLDVQTVVSLLRLYQYSNPAV
XP_001101017.1_1(190-298)      EGEGGHSAFQRILSKVREG-SLDVQTVVSLLRLYQDSNPAV
XP_001101280.1_1(190-298)      EGEGGHSAFQRILSKVREG-SLDVQTVVSLLRLYQDSNPAV
XP_001101193.1_1(237-345)      EGEGGHSAFQRILSKVREG-SLDVQTVVSLLRLYQDSNPAV
XP_544510.2_1(238-346)         EGEEGHSAFQRILDKVRDE-SLDVQTVVSLLRLYQDSNPAV
XP_614579.2_1(234-341)         EGE-GQSAFQRILDKVRSE-SLGVQTVVSLLRLYQDSNPAV
NP_937891.1_1(237-345)         TSEGGSSAFQRILDKVHDG-SLDVQVALSLVRLYQESTPAE
BAD90181.1_1(213-321)          TSEGGSSAFQRILDKVHDG-SLDVQVALSLVRLYQESTPAE
XP_342954.3_1(237-343)         EAEGGGSAFQRILDKVHDG-SLDVQVALSLMRLCQESTPAE
                                .    . .:.:* :: .  :* .:  :.*::  :   *
consensus/80%                  EGcss+pshpplLt+hpE-.oLsspslVpLLphhp.o.Psl
```

```
Secondary structure              HHHHHHHH
BAA32323.2_2(361-472)          GLLLEKLQKSATLPSTTVQPSP 112
Q9NUA8_2(347-458)              GLLLEKLQKSATLPSTTVQPSP 112
NP_055685.2_2(347-458)         GLLLEKLQKSATLPSTTVQPSP 112
CAI22041.1_2(347-453)          GLLLEKLQKSATLPSTT----- 107
XP_001164879.1_2(347-458)      GLLLEKLQKSATFPSATVQPSP 112
XP_001164989.1_2(300-411)      GLLLEKLQKSATFPSATVQPSP 112
XP_001164955.1_2(347-458)      GLLLEKLQKSATFPSATVQPSP 112
XP_001101017.1_2(300-411)      GLLLEKLQKLATLPGATVQPSP 112
XP_001101280.1_2(300-411)      GLLLEKLQKLATLPGATVQPSP 112
XP_001101193.1_2(347-458)      GLLLEKLQKLATLPGATVQPSP 112
NP_937891.1_2(337-448)         DLLLDNLQKGAGSAGTTGLARV 112
BAD90181.1_2(313-424)          DLLLDNLQKGAGSAGTTGLARV 112
XP_919018.2_1(78-189)          DLLLDNLQKGAGSAGTTGLARV 112
XP_342954.3_2(335-446)         GLLLENLQKVAESPGTTGLTRA 112
XP_544510.2_2(348-459)         GLMLENLQRLATWPSTTVQASP 112
XP_614579.2_2(343-454)         GLLLENLQKLATLPSTTAQANP 112
AAI14608.1_1(237-346)          GLLLEKLQKSATLPSTTVQPSP 110
XP_001165023.1_1(237-346)      GLLLEKLQKSATFPSATVQPSP 110
XP_001253042.1_1(234-343)      GLLLENLQKLATLPSTTAQANP 110
XP_001164989.1_1(190-298)      KTALLDRKPEDVDTVQPKGSTE 109
XP_001164955.1_1(237-345)      KTALLDRKPEDVDTVQPKGSTE 109
XP_001164879.1_1(237-345)      KTALLDRKPEDVDTVQPKGSTE 109
BAA32323.2_1(251-359)          KTALLDRKPEDVDTVQPKGSTE 109
Q9NUA8_1(237-345)              KTALLDRKPEDVDTVQPKGSTE 109
NP_055685.2_1(237-345)         KTALLDRKPEDVDTVQPKGSTE 109
CAI22041.1_1(237-345)          KTALLDRKPEDVDTVQPKGSTE 109
XP_001101017.1_1(190-298)      KTALLARKPEDVDTVQPKGSTE 109
XP_001101280.1_1(190-298)      KTALLARKPEDVDTVQPKGSTE 109
XP_001101193.1_1(237-345)      KTALLARKPEDVDTVQPKGSTE 109
XP_544510.2_1(238-346)         KAALLGRKPEGVEAVQPKGSTE 109
XP_614579.2_1(234-341)         KTALSDRKLEAVEAVQPKGSTE 108
NP_937891.1_1(237-345)         KVSQIQPEGSAGEGKTLSVLLL 109
BAD90181.1_1(213-321)          KVSQIQPEGSAGEGKTLSVLLL 109
XP_342954.3_1(237-343)         KVSQI--EGSAGEGKTLSVLLL 107
                                     :
consensus/80%                  thhL.php..ss.sstsh.sp.
```

The multiple sequence alignments corresponding to representative repeats and domains from various proteins along with their GENE or SWall identifiers. (a) GPA domain, (b) RxH domain, (c) GLG domain, (d) SAS domain, (e) WKRK domain, (f) FSS domain and (g) LLE domain. The numbers given in brackets indicate the start and end of amino acid residue positions corresponding to either the repeat or domain. The 82% consensus is labeled according to the alignment generated at the website www.bork.embl-heidelberg.de/Alignment/consensus.html**:** alcohol (o, ST); aliphatic (I, ILV); any (., ACDEFGHIKLMNPQRSTVWY); aromatic (a, FHWY); charged (c, DEHKR); hydrophobic (h, ACFGHIKLMRTVWY); negative (-, DE); polar (p, CDEHKNQRST); positive (+, HKR); small (s, ACDGNPSTV); tiny (u, AGS); turn-like (t, ACDEGHKNQRST). A capital letter indicates 82% conservation of corresponding amino acid residue. The secondary structure prediction indicated at the top was derived using the PROSITE program. Residues predicted with greater than 82% accuracy to form α helices are represented by 'H', β sheets are represented by 'E', loops are represented by 'L', coils are represented by 'C'.

Figure 6.2a: Multiple sequence alignment of 30 amino acid residue PGQY repeat.

```
Secondary structure                          HHHHHHHH
XP_059954.4_2(162-191)          LSLPFHGCLLDLGTCQAEPGQYCKEEVHIQ 30
Q5W0N0_1(86-115)                LSLPFHGCLLDLGTCQAEPGQYCKEEVHIQ 30
AAI30405.1_1(52-81)             LSLPFHGCLLDLGTCQAEPGQYCKEEVHIQ 30
XP_001147565.1_1(160-189)       LSLPFHGCLLDLGTCQAEPGQYCKEEVHVQ 30
XP_001253313.1_1(106-135)       LSIPFHGCLLDFGTCRTKPGQYCIKEVLIK 30
XP_001147565.1_2(119-148)       LGVILFGRLGDLGTCQTKPGQYWKEEVHIQ 30
Q5W0N0_2(45-74)                 LGVILFGRLGDLGTCQTKPGQYWKEEVHIQ 30
XP_059954.4_1(121-150)          LGVILFGRLGDLGTCQTKPGQYWKEEVHIQ 30
AAI30405.1_2(11-40)             LAPWLRPPFSDLGTCQTKPGQYWKEEVHIQ 30
                                *.  :   : *:***:::****  :** ::
consensus/80%                   Lul.haGpLhDLGTCQscPGQYhKEEVHIQ
```

Figure 6.2b: Multiple sequence alignment of 31 amino acid residue FYE repeat.

```
Secondary structure              HHHHHH H      HHHHHHHHHHHH
XP_856754.1(212-242)          LGGLQALVRE-VWLEKPQYDAAERGFYEAMFD 31
XP_532345.2(212-242)          LGGLQALVRE-VWLEKPQYDAAERGFYEAMFD 31
NP_083939.1(235-265)          LGSLQALVRE-VWLEKPRYDAAERGFYEALFD 31
NP_001013122.1(226-256)       LGSLQALVRE-VWLEKPRYDAAERGFYEALFD 31
AAP36729.1(222-252)           LGSLQALVRE-VWLEKPRYDAAERGFYEALFD 31
XP_594628.3(214-244)          LGSLQALVRE-VWLEKPRYDAAERGFYEALFD 31
AAH07847.1(222-252)           LGSLQALVRE-VWLEKPRYDAAERGFYEALFD 31
BAB14925.1(222-252)           LGSLQALVRE-VWLEKPRYDAAERGFYEALFD 31
NP_115754.2(222-252)          LGSLQALVRE-VWLEKPRYDAAERGFYEALFD 31
AAQ15199.1(222-252)           LGSLQALVRE-VWLEKPRYDAAERGFYEALFD 31
EAW82231.1(222-252)           LGSLQALVRE-VWLEKPRYDAAERGFYEALFD 31
XP_519999.2(221-251)          LGSLQALVRE-VWLEKPRYDAAERGFYEALFD 31
AAH00678.2(149-179)           LGSLQALVRE-VWLEKPRYDAAERGFYEALFD 31
BAE01260.1(244-274)           LGSLQALVRE-VWLEKPRYDAAERGFYEALFD 31
XP_001232628.1(257-287)       ASSLQALMSE-VWLEKPLYDGAEKSFYENMFD 31
CAI21007.1_2(281-312)         MTAADCLASERIWFDKPRYDEAERRFYEQMNG 32
CAI21006.1(1-32)              MTAADCLASERIWFDKPRYDEAERRFYEQMNG 32
NP_001025318.1(1-32)          MTAADCLASERIWFDKPRYDEAERRFYEQMNG 32
XP_688381.1(173-204)          MSGLQGLAQENIWFDKSRYDEAERCFYEGANG 32
XP_709090.1(1-32)             MSGLQGLAQENIWFDKSRYDEAERCFYEGANG 32
CAA59420.1(1-31)              MS-ASVIATEQVWLDKYKYDDAERQYYENLSG 31
CAF98101.1(240-271)           LPRIPVELLRDVWLEKPLYDRAEAVFYQNLYG 32
CAI21007.1_1(157-187)         LPCLSLPPMG-VWLQKPLFDKAEASFYQNLYN 31
                                        :*::*  :* **  :*:   .
consensus/80%                 hsulpsLspE.lWh-KPpYDtAERsFYEsh.s
```

Figure 6.2c: Multiple sequence alignment of 34 amino acid residue VHMM repeat.

```
Secondary structure                 HHHHHH              HH
AAH28606_3(170-203)           SLGVPQRGDLEDLEEHVPGQTVSEEATGVHMMQV 34
BAA34472_2(194-227)           SLGVPQRGDLEDLEEHVPGQTVSEEATGVHMMQV 34
NP_001072997_2(247-280)       SLGVPQRGDLEDLEEHVPGQTVSEEATGVHMMQV 34
AAD42863_2(170-203)           SLGVPQRGDLEDLEEHVPGQTVSEEATGVHMMQV 34
NP_775934_2(170-203)          SLGVPQRGDLEDLEEHVPGQTVSEEATGVHMMQV 34
BAF82207_2(170-203)           SLGVPQRGDLEDLEEHVPGQTVSEEATGVHMMQV 34
AAI01340(38-71)               SLGVPQRGDLEDLEEHLPGQTVSEEATGVHMMQV 34
NP_001072995(38-71)           SLGVPQRGDLEDLEEHLPGQTVSEEATGVHMMQV 34
NP_001072997_3(281-314)       DPATPAKSDLEDLEEHVPGQTVSEEATGVHMMQV 34
AAH28606_2(204-237)           DPATPAKSDLEDLEEHVPGQTVSEEATGVHMMQV 34
AAD42863_3(204-237)           DPATLAKSDLEDLEEHVPEQTVSEEATGVHMMQV 34
BAA34472_3(228-261)           DPATLAKSDLEDLEEHVPEQTVSEEATGVHMMQV 34
NP_775934_3(204-237)          DPATLAKSDLEDLEEHVPEQTVSEEATGVHMMQV 34
BAF82207_3(204-237)           DPATLAKSDLEDLEEHVPEQTVSEEATGVHMMQV 34
NP_001072997_1(181-214)       DTGIQTNGDLEDLEEHGPGQTVSEEATEVHMMEG 34
AAH28606_1(104-137)           DTGIQTNGDLEDLEEHGPGQTVSEEATEVHMMEG 34
AAD42863_1(104-137)           DAGTQTNGDLEDLEEHGPGQTVSEEATEVHMMEG 34
NP_775934_1(104-137)          DAGTQTNGDLEDLEEHGPGQTVSEEATEVHTMEG 34
BAA34472_1(128-161)           DAGTQTNGDLEDLEEHGPGQTVSEEATEVHMMEG 34
BAF82207_1(104-137)           DAGTQTNGDLEDLEEHGPGQTVSEEATEVHMMEE 34
                              . .   ..******** * ******** ** *:
consensus/80%                 s.us.tpuDLEDLEEHsPGQTVSEEATtVHMMps
```

Figure 6.2d: Multiple sequence alignment of 34 amino acid residue TQG repeat.

```
Secondary structure           CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC
NP_149020.2_1(478-511)        TIEEAPPTKKPLILKRKHATQGTMSHLKKPLILQ 34
BAF85143_1(478-511)           TIEEAPPTKKPLILKRKHATQGTMSHLKKPLILQ 34
EAW89921.1_1(478-511)         TIEEAPPTKKPLILKRKHATQGTMSHLKKPLILQ 34
Q8WWL7_1(478-511)             TIEEAPPTKKPLILKRKHATQGTMSHLKKPLILQ 34
CAC40024.1_1(478-511)         TIEEAPPTKKPLILKKKHATQGTMSHLKKPLILQ 34
Q8WWL7_2(575-608)             TTEETVLTKTSLSLQEKKITQGKMSHLKKPLVLQ 34
NP_149020.2_2(575-608)        TTEETVLTKTSLSLQEKKITQGKMSHLKKPLVLQ 34
BAF85143_2(575-608)           TTEETVLTKTSLSLQEKKITQGKMSHLKKPLVLQ 34
XP_521063.2_1(301-334)        TTEETVLTKTSLSLQEKKITQGKMSHLKKPLVLQ 34
EAW89921.1_2(575-608)         TTEETVLTKTSLSLQEKKITQGKMSHLKKPLVLQ 34
CAC40024.1_2(575-608)         TTEETVLTKTSLSLQEKKITQGEMSHLKKPLVLQ 34
NP_001005763.1_1(421-454)     TTKETNPTKKPLPIKKKCTIQGKMYYLKKPLVLQ 34
                              * :*:  **..* ::.*   ** * :*****:**
consensus/80%                 ThEEss.TKpsL.LpcK+hTQGpMSHLKKPLlLQ
```

Figure 6.2e: Multiple sequence alignment of 51 amino acid residue PES repeat.

```
Secondary structure          CCCCCCCCCCCCCCCCCCCCCCCCC CCCCCCCCCCCCCCC
NP_149020.2_1(955-1005)      PTYKEDTFLKTLLVPQVGTSPNVSS-TAPESITSKSSIATM
EAW89921.1_1(955-1005)       PTYKEDTFLKTLLVPQVGTSPNVSS-TAPESITSKSSIATM
CAC40024.1_1(955-1005)       PTYKEDTFLKTLLVPQVGTSPNVSS-TAPESITSKSSIATM
BAF85143_1(955-1005)         PTYKEDTFLKTLLVPQVGTSPNVSS-TAPESITSKSSIATM
Q8WWL7_1(955-1005)           PTYKEDTFLKTLLVPQVGTSPNVSS-TAPESITSKSSIATM
XP_521063.2_1(681-731)       PTYKEDTFLKTFLVPQVGTSPNVSS-TAPESITSKSSIATM
EAW89921.1_2(1037-1087)      PTCKEDTFLETFLIPQIGTSPYVFS-TTPESITEKSSIATM
BAF85143_2(1037-1087)        PTCKEDTFLETFLIPQIGTSPYVFS-TTPESITEKSSIATM
Q8WWL7_2(1037-1087)          PTCKEDTFLETFLIPQIGTSPYVFS-TTPESITEKSSIATM
NP_149020.2_2(1037-1087)     PTCKEDTFLETFLIPQIGTSPYVFS-TTPESITEKSSIATM
CAC40024.1_2(1037-1087)      PTCKEDTFLETFLIPQIGTSPYVFS-TTPESITEKSSIATM
XP_521063.2_2(763-814)       PTCKEDTFLETFLIPQIGTSPYVFSTTTPESITEKSSIATM
NP_001005763.1_1(986-1036)   PTQKEDTSLEDSLILQVETSSRVPS-TPPESRAGMSSVGKL
                             ** **** *:  *: *: **. * * *.*** :  **:..:
consensus/80%                PThKEDTFLcThLlPQlGTSP.V.S.TsPESITpKSSIATM


Secondary structure           CCCCCCCCCCC
NP_149020.2_1(955-1005)       TSVGKSGTINE 51
EAW89921.1_1(955-1005)        TSVGKSGTINE 51
CAC40024.1_1(955-1005)        TSVGKSGTINE 51
BAF85143_1(955-1005)          TSVGKSGTINE 51
Q8WWL7_1(955-1005)            TSVGKSGTINE 51
XP_521063.2_1(681-731)        TSVGKSGTINE 51
EAW89921.1_2(1037-1087)       TSVGKSRTTTE 51
BAF85143_2(1037-1087)         TSVGKSRTTTE 51
Q8WWL7_2(1037-1087)           TSVGKSRTTTE 51
NP_149020.2_2(1037-1087)      TSVGKSRTTTE 51
CAC40024.1_2(1037-1087)       TSVGKSRTTTE 51
XP_521063.2_2(763-814)        TSVGKSRTTTP 52
NP_001005763.1_1(986-1036)    STTSKSSVCES 51
                              ::..**  .
consensus/80%                 TSVGKStThsE
```

Figure 6.2f: Multiple sequence alignment of 34 amino acid residue HTQ repeat.

```
Secondary structure          CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC
EAL23925_3(92-125)           TPNPGQRRTHGHTQPRPAPDTWTHPTQASAGHTD 34
NP_001078865_3(92-125)       TPNPGQRRTHGHTQPRPAPDTWTHPTQASAGHTD 34
XP_374705_3(92-125)          TPNPGQRRTHGHTQPRPAPDTWTHPTQASAGHTD 34
XP_374705_2(58-91)           TPNPGQRRTHGHTQPRPAPDTRTHPTQASAGHTD 34
NP_001078865_2(58-91)        TPNPGQRRTHGHTQPRPAPDTRTHPTQASAGHTD 34
EAL23925_2(58-91)            TPNPGQRRTHGHTQPRPAPDTRTHPTQASAGHTD 34
XP_374705_4(126-159)         TPNPGQRRTHGHTQRRPVPDTRTHPIQASAGHTD 34
NP_001078865_4(126-159)      TPNPGQRRTHGHTQRRPVPDTRTHPIQASAGHTD 34
EAL23925_4(126-159)          TPNPGQRRTHGHTQRRPVPDTRTHPIQASAGHTD 34
EAL23925_1(4-37)             TDSPLERQTHRHTQRRPAPGTRTHPTQASAGHTD 34
NP_001078865_1(4-37)         TDSPLERQTHRHTQRRPAPGTRTHPTQASAGHTD 34
XP_374705_1(4-37)            TDSPLERQTHRHTQRRPAPGTRTHPTQASAGHTD 34
                             * .* :*:** *** **.*.* *** ********
consensus/80%                TssPhpRpTHtHTQ.RPsPsThTHPhQASAGHTD
```

Figure 6.2g: Multiple sequence alignment of 38 amino acid residue PTT repeat.

```
Secondary structure          CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC
EAL23926.1_5(379-416)        RVEQPTTGATSGAARKSPPQALRQAPLRTACHRSLSQG 38
EAL23925_5(625-662)          RVEQPTTGATSGAARKSPPQALRQAPLRTACHRSLSQG 38
XP_374705_4(625-662)         RVEQPTTGATSGAARKSPPQALRQAPPRTACHRSLSQG 38
EAL23926.1_1(34-71)          RPEQPTTDATSGADQKSPPQTLRQAPPERAHHRRYVRR 38
EAL23925_1(280-317)          RPEQPTTDATSGADQKSPPQTLRQAPPERAHHRRYVRR 38
NP_001078865_1(280-317)      RPEQPTTDATSGADQKSPPQTLRQAPPERAHHRRYVRR 38
XP_374705_1(280-317)         RPEQPTTDATSGADQKSPPQTLRQAPPERAHHRRYVRR 38
XP_374705_3(561-598)         RPEEPTTDATPGAARKSPPQTLRQAPPGRAHHRRYARR 38
NP_001078865_2(318-355)      RPEEPTTDATPGAARKSPPQTLRQAPPGRAHHRRYARR 38
EAL23925_2(318-355)          RPEEPTTDATPGAARKSPPQTLRQAPPGRAHHRRYARR 38
EAL23925_4(561-598)          RPEEPTTDATPGAARKSPPQTLRQAPPGRAHHRRYARR 38
EAL23926.1_2(72-110)         RPEEPTTDATPGAARKSPPQTLRQAPPGRAHHRRYARR 38
EAL23926.1_4(315-352)        RPEEPTTDATPGAARKSPPQTLRQAPPGRAHHRRYARR 38
XP_374705_5(318-355)         RPEEPTTDATPGAARKSPPQTLRQAPTGRAHHRRYARR 38
EAL23926.1_3(277-314)        RPEQPTTDATPGAARKSPPQTLRQAPPGRAHHRRYARR 38
EAL23925_3(523-560)          RPEQPTTDATPGAARKSPPQTLRQAPPGRAHHRRYARR 38
XP_374705_2(523-560)         RPEQPTTDATPGAARKSPPQTLRQAPTGRAHHRRYVRR 38
XP_374705_6(356-393)         RPEEPTTDATPGAAPNSPPQTVSQAPAWNSPPQMLRQT 38
EAL23926.1_6(111-147)        RPEEPTTDATPGAAPNSPPQTVSQAPAWNSPPQMLRQT 38
NP_001078865_3(356-393)      RPEEPTTDATPGAAPNSPPQTVSQAPAWNSPPQMLRQT 38
EAL23925_6(356-393)          RPEEPTTDATPGAAPNSPPQTVSQAPAWNSPPQMLRQT 38
                             * *:***.**.**  :****:: ***   :  :   :
consensus/80%                RPEpPTTDATsGAApKSPPQTLRQAPstpApHRphspp
```

Figure 6.2h: Multiple sequence alignment of 34 amino acid residue FSQ repeat.

```
Secondary structure          CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC
NP_001005333.1_1(126-159)    AAARPKSAFKVQNATTK-GPNGVYDFSQAHNAKDV 34
NP_008917.3_1(70-103)        AAARPKSAFKVQNATTK-GPNGVYDFSQAHNAKDV 34
AAG09704.1_1(70-103)         AAARPKSAFKVQNATTK-GPNGVYDFSQAHNAKDV 34
EAW62896.1_1(70-103)         AAARPKSAFKVQNATTK-GPNGVYDFSQAHNAKDV 34
BAB84918.1_1(84-117)         AAARPKSAFKVQNATTK-GPNGVYDFSQAHNAKDV 34
BAD51991.1_1(70-103)         AAARPKSAFKVQNATTK-GPNGVYDFSQAHNAKDM 34
BAD90336.1_1(122-152)        AAARPKTGFKAQNATTK-GPN---DYSQARNAKEM 31
NP_062765.1_1(75-105)        AAARPKTGFKAQNATTK-GPN---DYSQARNAKEM 31
AAK01203.1_1(75-105)         AAARPKTGFKAQNATTK-GPN---DYSQARNAKEM 31
AAH31461.1_1(75-105)         AAARPKTGFKAQNATTK-GPN---DYSQARNAKEM 31
BAE27491.1_1(75-105)         AAARPKTGFKAQNATTK-GPN---DYSQARNAKEM 31
AAH16438.1_1(75-105)         AAARPKTGFKAQNATTK-GPN---DYSQARNAKEM 31
Q9ES73_1(75-105)             AAARPKTGFKAQNTTTK-GPN---DYSQARNAKEM 31
NP_445861.1_1(75-105)        AAARPKTGFKAQNTTTK-GPN---DYSQARNAKEM 31
BAE22540.1_1(75-105)         AAARPKTGFKVQNATTK-GPN---DYSQARNAKEM 31
NP_001001860.1_1(75-108)     AATKPKTAFKAQNATTK-GPNAAYDFSQALNAKEI 34
NP_001039590.1_1(75-108)     TATKPKTAFKVQNATTK-GPNAAYDFSQAFNAKET 34
NP_008917.3_3(138-172)       AANKSEMAFKAQNATTKVGPNATYNFSQSLNANDL 35
BAB84918.1_3(152-186)        AANKSEMAFKAQNATTKVGPNATYNFSQSLNANDL 35
AAG09704.1_3(138-172)        AANKSEMAFKAQNATTKVGPNATYNFSQSLNANDL 35
EAW62896.1_3(138-172)        AANKSEMAFKAQNATTKVGPNATYNFSQSLNANDL 35
NP_001005333.1_3(194-228)    AANKSEMAFKAQNATTKVGPNATYNFSQSLNANDL 35
XP_538044.2_2(36-70)         TANKSEMAFKAQNATTKVGPNATYNFSQSLSATEM 35
NP_001039590.1_3(143-177)    TANKSDMAFKAQNTTTKVVPNAAYNFSQSLNASEM 35
BAE22540.1_3(139-172)        SAKKSEMAFKGQNS-TKAGPGTTYNFPQSPSANEM 34
AAH31461.1_3(139-172)        SAKKSEMAFKGQNS-TKAGPGTTYNFPQSPSANEM 34
BAE27491.1_3(139-172)        SAKKSEMAFKGQNS-TKAGPGTTYNFPQSPSANEM 34
AAH16438.1_3(139-172)        SAKKSEMAFKGQNS-TKAGPGTTYNFPQSPSANEM 34
BAD90336.1_3(186-219)        SAKKSEMAFKGQNS-TKAGPGTTYNFPQSPSANEM 34
NP_062765.1_3(139-171)        SAKKSEMAFKGQNS-TKAGPGTTYNFPQSPSANEM 34
AAK01203.1_3(139-172)        SAKKSEMAFKGQNS-TKAGPGTTYNFPQSPSANEM 34
Q9ES73_3(140-174)            AAKKSEMAFKGQNTTTKAGPSATYNFTQSPSANEM 35
NP_445861.1_3(140-174)       AAKKSEMAFKGQNTTTKAGPSATYNFTQSPSANEM 35
NP_001001860.1_3(142-176)    LTAKPEMAFKAQNATTKVGPNATYNFSPSLNANEM 35
EAW62896.1_2(104-137)        PNTQPKAAFKSQNATPK-GPNAAYDFSQAATTGEL 34
BAB84918.1_2(118-151)        PNTQPKAAFKSQNATPK-GPNAAYDFSQAATTGEL 34
NP_008917.3_2(104-137)       PNTQPKAAFKSQNATPK-GPNAAYDFSQAATTGEL 34
NP_001005333.1_2(160-193)    PNTQPKAAFKSQNATPK-GPNAAYDFSQAATTGEL 34
AAG09704.1_2(104-137)        PNTQPKAAFKSQNATSK-GPNAAYDFSQAATTGEL 34
NP_001039590.1_2(109-142)    PNILPTAHFKSQNAPAK-GPNAAYDFSQAAPTGEL 34
XP_538044.2_1(2-35)          PNVPPKAAFKSQNATLK-GPNAAYDFSQAATTSEL 34
NP_001001860.1_2(109-142)    PSTPPTVAFKAPNAPSK-GPNAAYDFSQAATTSEL 34
AAH31461.1_2(106-139)        PKNQSKAAFKSQNGTPK-GSHAASDFSQAAPTGKS 34
AAH16438.1_2(106-139)        PKNQSKAAFKSQNGTPK-GPHAASDFSQAAPTGKS 34
BAE27491.1_2(106-139)        PKNQSKAAFKSQNGTPK-GPHAASDFSQAAPTGKS 34
BAD90336.1_2(153-186)        PKNQSKAAFKSQNGTPK-GPHAASDFSQAAPTGKS 34
NP_062765.1_2(106-139)       PKNQSKAAFKSQNGTPK-GPHAASDFSQAAPTGKS 34
BAE22540.1_2(106-139)        PKNQSKAAFKSQNGTPK-GPHAASDFSQAAPTGKS 34
AAK01203.1_2(106-139)        PKNQSKAAFKSQNGTPK-GPHAASDFSQAAPTGKS 34
Q9ES73_2(106-139)            PKNQPKVAFKSQNATSK-GPHAASDFSHAASTGKS 34
NP_445861.1_2(106-139)       PKNQPKVAFKSQNATSK-GPHAASDFSHAASTGKS 34
                                .  **  *  *  .   ::. :  : .
consensus/80%                sspspschAFKuQNuTsK.GPsss.sFSQuhsst-h
```

Figure 6.2i: Multiple sequence alignment of 36 amino acid residue PEG repeat.

```
Secondary structure          CCCCCCCCCCCCC   CCCCCCCCCCCCCCCCCCCCCCC
AAC24227_2(58-93)            SEGEDSSDPLQRP---PEGKDSQSPLQIPQSSPEGDDTQ 36
AAC18837_2(58-93)            SEGEDSSDPLQRP---PEGKDSQSPLQIPQSSPEGDDTQ 36
NP_005453_2(58-93)           SEGEDSSDPLQRP---PEGKDSQSPLQIPQSSPEGDDTQ 36
XP_942006_1(48-83)           SEGEDSSDPLQRP---PEGKDSQSPLQIPQSSPEGDDTQ 36
NP_005453_3(668-705)         PEGMHSQSPLQSPESAPEGEDSLSPLQIPQSPLEGEDSL 39
AAC24227_3(667-705)          PEGMHSQSPLQSPESAPEGEDSLSPLQIPQSPLEGEDSL 39
AAC18837_3(667-705)          PEGMHSQSPLQSPESAPEGEDSLSPLQIPQSPLEGEDSL 39
XP_001126506.1(326-364)      PEGMHSQSPLQSPESAPEGEDSLSPLQIPQSPLEGEDSL 39
NP_005453_1(12-48)           PSLLQSSS--ESPQSCPEGEDSQSPLQIPQSSPESDDTL 37
XP_942006_2(2-38)            PSLLQSSS--ESPQSCPEGEDSQSPLQIPQSSPESDDTL 37
AAC24227_1(12-48)            PSLLQSSS--ESPQSCPEGEDSQSPLQIPQSSPESDDTL 37
AAC18837_1(12-48)            PSLLQSSS--ESPQSYPEGEDSQSPLQIPQSSPESDDTL 37
                             ..  .*..  : *   ***:** ********. *.:*:
consensus/80%                sph.pSps..ppP...PEGcDS.SPLQIPQSs.Eu-Do.
```

Figure 6.2j: Multiple sequence alignment of 42 amino acid residue SSC repeat.

```
Secondary structure                HHHHH
XP_001127353(151-192)        SQEPLRAQLLPPSGLYR-PSTCLRAAFPGLAFAALHPLQALDF 42
XP_001104785.1_1(10-52)      VRRPLPAWLLPPGRLVRRPGACPRAAFPGLDFAALHPLQALNL 43
EAW83660.1(143-184)          FQQLLHTQLLPPSGLFR-PSSCFTRAFPGPTFVSWQPSLARFL 42
XP_530161.2(106-147)         FQQLLHTQLLPPSGLFR-PSSCFTRAFPGPTFVSWQPSLARFL 42
EAW83662.1(152-193)          FQQLLHTQLLPPSGLFR-PSSCFTRAFPGPTFVSWQPSLARFL 42
XP_934335_2(120-161)         FQQLLHTQLLPPSGLFR-PSSCFTRAFPGPTFVSWQPSLARFL 42
AAI12925_1(13-54)            FQQLLHTQLLPPSGLFR-PSSCFTRAFPGPTFVSWQPSLARFL 42
EAW61038.1_1(159-200)        FQQLLHTQLLPPSGLFR-PSSCLTRAFPGPTFVSWQPSLATFL 42
EAW69824.1_2(105-146)        FQQLLHTQLLPPSGLFR-PSSCFTRAFPGSTFVSWQPFLARFL 42
EAW69822_1(150-191)          FQQLLHTQLLPPSGLFR-PSSCFTRAFPGSTFVSWQPFLARFL 42
EAW69823_2(143-184)          FQQLLHTQLLPPSGLFR-PSSCFTRAFPGSTFVSWQPFLARFL 42
XP_934335_1(45-85)           LPAFSPGPELSQVNLTR-PSSCFFAASPGPAPASWWPLQAQPL 42
EAW69822_3(75-116)           LPAFSPGPELSQVDLTR-PSSCFFAASPGPAPASWWPLQAQPL 42
EAW61038.1_3(84-125)         LPAFSPGPELSQVDLTR-PSSCFLAASPGPAPASWQPLQAQPL 42
EAW69822_2(194-235)          SQQPRQAQVLPHTGLST-SSSCLTVASPGPTPVPGRHLRAQNL 42
EAW69823_1(187-228)          SQQPRQAQVLPHTGLST-SSSCLTVASPGPTPVPGRHLRAQNL 42
EAW69824.1_1(149-190)        SQQPRQAQVLPHTGLST-SSSCLTVASPGPTPVPGRHLRAQNL 42
EAW61038.1_2(203-244)        SQQPRQAQVLPHTGLST-SSSCLTVASPGPAPVPGRHLQAQNF 42
NP_001004321.2(59-100)       SSQALRAHLLPPGGLYS-SSTGWRTASAGPALASQGPLQAQLL 42
BAC87609.1(59-100)           SSQALRAHLLPPGGLYS-SSTGWRTASAGPALASQGPLQAQLL 42
                                   *.   *   ..:    * .*   ..     *  :
consensus/80%                .pp.hpsplLP.sGLhp.sSSChptA.PGss.suhpP..Ap.L
```

Figure 6.2k: Multiple sequence alignment of 42 amino acid residue YCL repeat.

```
Secondary structure                                E HHHHHHHHHHHHH
XP_516170.2_2(556-599)          VSPSKTLSVLDKDVPGHGRNRYDEIKEEFDKLHQKYCLKSPGDP 44
XP_001151551.1_2(557-600)       VSPSKTLSVLDKDVPGHGRNRYDEIKEEFDKLHQKYCLKSPGDP 44
BAC11221.1_2(556-599)           VSPSKTLSVPDKEVPGHGRNRYDEIKEEFDKLHQKYCLKSPGDP 44
EAW71064.1_2(556-599)           VSPSKTLSVPDKEVPGHGRNRYDEIKEEFDKLHQKYCLKSPGDP 44
NP_060880.3_2(556-599)          VSPSKTLSVPDKEVPGHGRNRYDEIKEEFDKLHQKYCLKSPGDP 44
EAW71065.1_2(332-375)           VSPSKTLSVPDKEVPGHGRNRYDEIKEEFDKLHQKYCLKSPGDP 44
XP_001110520.1_2(678-721)       VSPIKTLSGPDKEVLGHRRNRYDEIKEEFDKLHQKYCLRSPEQM 44
BAE02517.1_2(285-328)           VSPIKTLSGPDKEVLGHRRNRYDEIKEEFDKLHQKYCLRSPEQM 44
XP_874813.2_2(661-704)          VSFSKAISVPRVQPLGFARDRYDDIKEKFDKLHRQYCQKSPQER 44
XP_516170.2_1(411-452)          ISPVKIVSRPTIRQ-GHGENRQREIEIRFDQLHREYCL-SPRNQ 42
BAC11221.1_1(411-452)           ISPVKIVSRPTIRQ-GHGENRQREIEIRFDQLHREYCL-SPRNQ 42
NP_060880.3_1(411-452)          ISPVKIVSRPTIRQ-GHGENRQREIEIRFDQLHREYCL-SPRNQ 42
EAW71065.1_1(187-228)           ISPVKIVSRPTIRQ-GHGENRQREIEIRFDQLHREYCL-SPRNQ 42
XP_001151551.1_1(412-453)       ISPVKIVSRPTIRQ-GHGENRQREIEIRFDQLHREYCL-SPRNQ 42
EAW71064.1_1(411-452)           ISPVKIVSRPTIRQ-GHGENRQREIEIRFDQLHREYCL-SPRNQ 42
XP_001110520.1_1(534-575)       ISPVKIVSRPRIQQ-GHRENRQREIETRFDQLHREYCL-SPRNQ 42
BAE02517.1_1(141-182)           ISPVKIVSRPRIQQ-GHRENRQREIETRFDQLHREYCL-SPRNQ 42
XP_874813.2_1(517-558)          ISPVKIASRPRILR-GQAGSRFKEIEIKFDKLHQEYCL-SPGKQ 42
XP_001065693.1_1(394-435)       ISPVKMAPRLRVPP-GQVENWYREVKIKFEKLHQECCP-SPEKR 42
XP_237403.4_1(394-435)          ISPVKMAPRLRVPP-GQVENWYREVKIKFEKLHQECCP-SPEKR 42
NP_766093.1_1(311-352)          ISPVKVVPRPRMLP-SQVEKWYREIKIKFDKLHQEYCL-SSGKQ 42
XP_001005388.1_1(281-322)       ISPVKVVPRPRMLP-SQVEKWYREIKIKFDKLHQEYCL-SSGKQ 42
NP_941054.1_1(387-428)          ISPVKVVPRPRMLP-SQVEKWYREIKIKFDKLHQEYCL-SSGKQ 42
BAD36742.1_1(370-411)           ISPVKVVPRPRMLP-SQVEKWYREIKIKFDKLHQEYCL-SSGKQ 42
BAC27950.1_1(108-149)           ISPVKVVPRPRMLP-SPVEKWYREIKIKFDKLHQEYCL-SSGKQ 42
AAH62125.1_1(121-162)           ISPVKVVPRPRMLP-SPVEKWYREIKIKFDKLHQEYCL-SSGKQ 42
BAD36742.1_2(512-554)           VWPSTAISAPSIGSPGCGKDNYDELKKEFNRLYQKYCL-SPQRA 43
NP_941054.1_2(529-571)          VWPSTAISAPSIGSPGCGKDNYYELKKEFNRLYQKYCL-SPQRA 43
                                :   .  .       .   .   ::: .*::*::: *  *.
consensus/80%                   lSPsKhlshPph...upscph.cEIc.cFDpLHpcYCL.Sstp.
```

Figure 6.2l: Multiple sequence alignment of 43 amino acid residue VSR repeat.

```
Secondary structure                        HHH
XP_374142.3(408-450)          RCPSPQNLSPWDRTTRRVSSPLFPE-ASSEWENQNPAVEETVSR 43
XP_001083163.1(407-449)       RCPSPQNLSPWDRTSQRVSSPLFPE-ASSAWENQNLAVEETVSR 43
XP_001502524(409-451)         HCPSPQNLSPWNRAIRRANSPSFPK-ASSAWENQDPAVKETVSR 43
XP_896829.1(465-508)          PNGSPHNPPQPSKIVWRRSSPAFPGGTSSAWENRNASVEEVVSR 44
BAE26839.1(400-443)           PNGSPHNPPQPSKIVWRRSSPAFPGGTSSAWENRNASVEEVVSR 44
XP_001069975.1(410-452)       RCPSPQNLPQPHKIVWRTSSPAFPG-ASSAWEDRNASVEEVVSR 43
                                 **:* .   :   * .** **  :** **::: :*:*.***
consensus/80%                 .ssSPpN.s..s+hshRhSSPhFPt.sSSAWENpNsuVEEsVSR
```

Figure 6.2m: Multiple sequence alignment of 54 amino acid residue ALPG repeat.

```
Secondary structure           LLLLLLLLLLLLLLLLLLLLLLLLLLLLLL
XP_896829.1(66-118)        LAPLALPGLSRRLP-APAGRQDSSGSSGSYHTAPGSPE
BAE26839.1(1-53)           LAPLALPGLSRRLP-APAGRQDSSGSSGSYHTAPGSPE
XP_001069975.1(5-57)       LAPLALPGLSRRLP-ASAGRQDSSGSSGSYHTAPGSPE
XP_001083163.1(5-58)       LAPPALPGIPRQLPTAPARRQDSSGSSGSYYTAPGSPE
XP_374142.3(5-58)          LAPPALPGIPRQLPTAPARRQDSSGSSGSYYTAPGSPE
XP_001502524(5-58)         LSPPALPGLPGRLPAAPARRQDSSGSSGSYYTAPGSPE
                           *:* ****:. :** *.* ***********:*******
consensus/80%              LAP.ALPGlsRpLP.APAtRQDSSGSSGSYaTAPGSPE


Secondary structure       LLLLLLLLLL
XP_896829.1(66-118)       PPDVGPDTEGRGTWLW 53
BAE26839.1(1-53)          PPDVGPDTEGRGTWLW 53
XP_001069975.1(5-57)      PPDVGPDTEGQGTWLW 53
XP_001083163.1(5-58)      PPDVGPDAEGPANWPW 54
XP_374142.3(5-58)         PPDVGPDAKGPANWPW 54
XP_001502524(5-58)        PPDVGPDAEGRANWPG 54
                          *******::* ..*
consensus/80%             PPDVGPDsEG.usW.W
```

Figure 6.2n: Multiple sequence alignment of 43 amino acid residue SVT repeat.

```
Secondary structure          CCCCCCCCCCCCCCCCCCCCCCCCCCCCC    CCCCCCCCCC
XP_499019.3_1(39-81)         PPQCQPRPSFPLPRSLLSVTTAIVPSATLPPQCHHGHRSLCHA 43
XP_499019.3_3(125-167)       PSSVSPRPSFPLPRSLLSVTTAIVPSATLPPQCHHGHRSLCHA 43
XP_499019.3_2(82-124)        PSSVSPRPSFPLPRSLLSVTTAIVPSATLPPQCHHGHRSLCHA 43
XP_499019.3_5(240-282)       PSSVSPWPSFPLPRSLLSVTMAIVPSATLPPQCHHGHRSLCHA 43
XP_499019.3_4(168-210)       PSSVSPRPSFPLPRSLLSVTTAIVPSATLPPQCHHGHRSFCHA 43
XP_499019.3_6(312-354)       PSSVSPRPSFPLPRSLLSVTTAIIPSATLPPQCHHGHHSTSRF 43
                             *.. .* ************ **:*************:* .:
consensus/80%                PSSVSPRPSFPLPRSLLSVTTAIVPSATLPPQCHHGHRShCHA
```

Figure 6.2o: Multiple sequence alignment of 46 amino acid residue CDxD repeat.

```
Secondary structure                    HHHHHHHHHHH          EEE  EE
AAD50912.2_1(64-109)             IVFSASGANLTDAHLFCLLACDRDLCCD--GFV--LTQVQ
EAW92157.1_1(1708-1753)          IVFSASGANLTDAHLFCLLACDRDLCCD--GFV--LTQVQ
NP_003226.4_1(1708-1753)         IVFSASGANLTDAHLFCLLACDRDLCCD--GFV--LTQVQ
BAD92396.1_1(514-559)            IVFSASGANLTDAHLFCLLACDRDLCCD--GFV--LTQVQ
AAC51924.1_1(1708-1753)          IVFSASGANLTDAHLFCLLACDRDLCCD--GFV--LTQVQ
CAA29104.1_1(1707-1752)          IVFSASGANLTDAHLFCLLACDRDLCCD--GFV--LTQVQ
P01266_1(1708-1753)              IVFSASGANLTDAHLFCLLACDRDLCCD--GFV--LTQVQ
AAB53204.1_1(1707-1752)          VVFSASGANLTDTHIYCLLACDNDSCCD--GFI--ITQVK
O08710_1(1707-1752)              VVFSASGANLTDTHIYCLLACDNDSCCD--GFI--ITQVK
AAC32269.1_1(1707-1752)          VVFSASGANLTDTHTYCLLACDNDSCCD--GFI--ITQVK
NP_033401.2_1(1707-1752)         VVFSASGANLTDTHTYCLLACDNDSCCD--GFI--ITQVK
AAC32268.1_1(1707-1752)          VVFSASGANLTDTHTYCLLACDNDSCCD--GFI--ITQVK
NP_112250.1_1(1707-1752)         VVFSALGTNLTDTHLFCLLACDQDSCCD--GFI--VTQVK
AAF34909.1_1(1707-1752)          VVFSALGTNLTDTHLFCLLACDQDSKSD--GFI--VTQVK
NP_776308.1_1(1710-1755)         VTFSASGASLAEVHLFCLLACDHDSCCD--GFI--LVQVQ
NP_001041569.1_1(1709-1754)      VIFPASGADLTAAHLFCLLACDRDSCCD--GFI--LAQLQ
NP_003226.4_2(1980-2028)         NKVPMSEKSISNGFFECERRCDADPCCTGFGFLN-VSQLK
P01266_2(1980-2028)              NKVPMSEKSISNGFFECERRCDADPCCTGFGFLN-VSQLK
AAD50912.2_2(336-384)            NKVPMSEKSISNGFFECERRCDADPCCTGFGFLN-VSQLK
EAW92157.1_2(1980-2028)          NKVPMSEKSISNGFFECERRCDADPCCTGFGFLN-VSQLK
CAA29104.1_2(1979-2027)          NKVPMSEKSISNGFFECERRCDADPCCTGFGFLN-VSQLK
BAD92396.1_2(786-834)            NKVPMSEKSISNGFFECERRCDADPCCTGFGFLN-VSQLK
AAC51924.1_2(1980-2028)          NKVPMSEKSISNGFFECERRCDADPCCTGFGFLN-VSQLK
AAC32268.1_2(1977-2025)          DKVPMSGKLISNGFFECERLCDRDPCCTGFGFLN-VSQLQ
NP_033401.2_2(1977-2025)         DKVPMSGKLISNGFFECERLCDRDPCCTGFGFLN-VSQLQ
AAC32269.1_2(1977-2025)          DKVPMSGKLISNGFFECERLCDRDPCCTGFGFLN-VSQLQ
AAB53204.1_2(1977-2025)          DKVPMSGKLISNGFFECERLCDRDPCCTGFGFLN-VSQLQ
O08710_2(1977-2025)              DKVPMSGKLISNGFFECERLCDRDPCCTGFGFLN-VSQLQ
AAF34909.1_2(1979-2027)          DRIPMSEKLISNGFFECERLCDRDPCCTGFGFLN-VSQMQ
CAA26183.1_1(178-226)            DRIPMSEKLISNGFFECERLCDRDPCCTGFGFLN-VSQMQ
NP_112250.1_2(1979-2027)         DRIPMSEKLISNGFFECERLCDRDPCCTGFGFLN-VSQMQ
CAF89701.1_1(1086-1133)          QVFSSRETSLSDLHRFCQDICRHDTCCH--GYIINQNSFK
                                   ..     ::  .  *   *  *  .   *::    ..:
consensus/80%                    .hhshStt.lossah.C.hhCDtD.CCs..GFl..loQlp
```

```
Secondary structure                EEEEE
AAD50912.2_1(64-109)            GGAIICGLLS 46
EAW92157.1_1(1708-1753)         GGAIICGLLS 46
NP_003226.4_1(1708-1753)        GGAIICGLLS 46
BAD92396.1_1(514-559)           GGAIICGLLS 46
AAC51924.1_1(1708-1753)         GGAIICGLLS 46
CAA29104.1_1(1707-1752)         GGAIICGLLS 46
P01266_1(1708-1753)             GGAIICGLLS 46
AAB53204.1_1(1707-1752)         GGPTICGLLS 46
O08710_1(1707-1752)             GGPTICGLLS 46
AAC32269.1_1(1707-1752)         GGPTICGLLS 46
NP_033401.2_1(1707-1752)        GGPTICGLLS 46
AAC32268.1_1(1707-1752)         GGPTICGLLS 46
NP_112250.1_1(1707-1752)        EGPTICGLLS 46
AAF34909.1_1(1707-1752)         EGPTICGLLS 46
NP_776308.1_1(1710-1755)        GGPLLCGLLS 46
NP_001041569.1_1(1709-1754)     GGPVICGLLS 46
NP_003226.4_2(1980-2028)        GGEVTCLTLN 49
P01266_2(1980-2028)             GGEVTCLTLN 49
AAD50912.2_2(336-384)           GGEVTCLTLN 49
EAW92157.1_2(1980-2028)         GGEVTCLTLN 49
CAA29104.1_2(1979-2027)         GGEVTCLTLN 49
BAD92396.1_2(786-834)           GGEVTCLTLN 49
AAC51924.1_2(1980-2028)         GGEVTCLTLN 49
AAC32268.1_2(1977-2025)         GGEVTCLTLN 49
NP_033401.2_2(1977-2025)        GGEVTCLTLN 49
AAC32269.1_2(1977-2025)         GGEVTCLTLN 49
AAB53204.1_2(1977-2025)         GGEVTCLTLN 49
O08710_2(1977-2025)             GGEVTCLTLN 49
AAF34909.1_2(1979-2027)         GGEMTCLTLN 49
CAA26183.1_1(178-226)           GGEMTCLTLN 49
NP_112250.1_2(1979-2027)        GGEMTCLTLN 49
CAF89701.1_1(1086-1133)         SGSLFCGWLG 48
                                 *   *  *.
consensus/80%                   GG.hhChhLs
```

Figure 6.2p: Multiple sequence alignment of 50 amino acid residue GGF repeat.

```
Secondary structure          CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC
XP_941683.2_1(116-165)       QDAGGSKSEDTPPGGFFYGSSSSGDSDKKKPLFSF
XP_497341.3_1(116-165)       QDAGGSKSEDTPPGGFFYGSSSSGDSDKKKPLFSF
XP_941683.2_3(66-115)        QDAGSPKSEDTPAGGFFNSSSSSGDSDRTKPFFSL
XP_497341.3_3(66-115)        QDAGSPKSEDTPAGGFFNSSSSSGDSDRTKPFFSL
XP_941683.2_2(17-65)         KDVEALLYRQK-SGGFSYGSSSSGDLDRKKPLFSL
XP_497341.3_2(17-65)         KDVEALLYRQK-SGGFSYGSSSSGDLDRKKPLFSL
                             :*. .   .:. .***  .****** *:.**:**:
consensus/80%                pDstu.h.cpp.sGGF..uSSSSGD.D+pKPhFSh


Secondary structure          CCCCCCCCCCCCCCCC
XP_941683.2_1(116-165)       EFGATGEDEDKSRER 50
XP_497341.3_1(116-165)       EFGATGEDEDKSRER 50
XP_941683.2_3(66-115)        GLGAPGKAEDKSGDS 50
XP_497341.3_3(66-115)        GLGAPGKAEDKSGDS 50
XP_941683.2_2(17-65)         EFGSPGETEDKSRQR 49
XP_497341.3_2(17-65)         EFGSPGETEDKSRQR 49
                              :*:.*: **** :
consensus/80%                thGusGcsEDKStpp
```

Figure 6.2q: Multiple sequence alignment of 52 amino acid residue NYS repeat.

```
Secondary strucuture     CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC
NP_048536.2_3(256-307)   LGMSFNRNRNYSRPIGPNPRENMSGSAGLFGQRPFDPQINYSAPIGPAQGPG 52
NP_048536.2_4(308-359)   LGMSFNRNRNYSAPIGPNPRENMSGSAGLFGQRPYNPKINYSAPIGPEPMPE 52
NP_048536.2_2(204-255)   LGMSFNRNRNYSVPIGPLPRQNMSGSAGLFGQRPFDQQINYSAPIGPAQGPG 52
NP_048536.2_1(152-203)   LGGSYDPLINYSAPIGPLPKQSASGSAGLFKNRPFDQQINYSQPIGPAQGPG 52
                         ** *::   *** **** *::. ******* :**:: :**** ****   *
consensus/80%            LGhSas..hNYShPIGP.P+pshSGSAGLFtpRPas.pINYStPIGPt.hPt
```

Figure 6.2r: Multiple sequence alignment of 52 amino acid residue RPE repeat.

```
Secondary structure                                EEE
BAA20760.2(1242-1294)    DLISSPGKKGAAHPDPSKTSVDTGQVSRPENPSQPASP
O15018(1234-1286)        DLISSPGKKGAAHPDPSKTSVDTGQVSRPENPSQPASP
XP_526957.2(837-889)     DLISSPGKKGAAHPDPSKTSVDTGQVSRPENPSQPASP
AAK07661.1(1036-1088)    DLISSPGKKGAAHPDPSKTSVDTGQVSRPENPSQPASP
NP_835260.2(1234-1286)   DLISSPGKKGAAHPDPSKTSVDTGQVSRPENPSQPASP
EAX10777.1(1060-1112)    DLISSPGKKGAAHPDPSKTSVDTGQVSRPENPSQPASP
XP_871254.2(1224-1277)   DPLPSPGQKEAAHPDPSQTSVDTEPARRPEDPGGPESP
XP_536512.2(1107-1159)   GLDGPPGQKGAAHPDPGEPSADTGHARRPEDPGKPVSL
                         .   .**:* ******.:.*.**  . ***:*. * *
consensus/80%            DLluSPGpKGAAHPDPScTSVDTGpspRPEsPupPsSP

Secondary structure
BAA20760.2(1242-1294)    RVAKCK-ARSPVRLPH 53
O15018(1234-1286)        RVAKCK-ARSPVRLPH 53
XP_526957.2(837-889)     RVAKCK-VRSPVRLPH 53
AAK07661.1(1036-1088)    RVTKCK-ARSPVRLPH 53
NP_835260.2(1234-1286)   RVTKCK-ARSPVRLPH 53
EAX10777.1(1060-1112)    RVTKCK-ARSPVRLPH 53
XP_871254.2(1224-1277)   RIPKSEDSTSPGTMAM 54
XP_536512.2(1107-1159)   GGSESE-DRGQARLAQ 53
                           .:.:   .   :.
consensus/80%            RlsKsc.sRSPsRLsp
```

The multiple sequence alignments corresponding to representative repeats and domains from various proteins along with their GENE or SWall identifiers. (a) PGQY repeat, (b) FYE repeat, (c) VHMM repeat, (d) TQG repeat, (e) PES repeat, (f) HTQ repeat, (g) PTT repeat, (h) FSQ repeat, (i) PEG repeat, (j) SSC repeat, (k) YCL repeat, (l) VSR repeat, (m) ALPG repeat, (n) SVT repeat, (o) CDxD repeat, (p) GGF repeat, (q) NYS repeat and (r) RPE repeat. The numbers given in brackets indicate the start and end of amino acid residue positions corresponding to either the repeat or domain. The 82% consensus is labeled according to the alignment generated at the website www.bork.embl-heidelberg.de/Alignment/consensus.html**:** alcohol (o, ST); aliphatic (I, ILV); any (., ACDEFGHIKLMNPQRSTVWY); aromatic (a, FHWY); charged (c, DEHKR); hydrophobic (h, ACFGHIKLMRTVWY); negative (-, DE); polar (p, CDEHKNQRST); positive (+, HKR); small (s, ACDGNPSTV); tiny (u, AGS); turn-like (t, ACDEGHKNQRST). A capital letter indicates 82% conservation of corresponding amino acid residue. The secondary structure prediction indicated at the top was derived using the PROSITE program. Residues predicted with greater than 82% accuracy to form α helices are represented by 'H', β sheets are represented by 'E', loops are represented by 'L', coils are represented by 'C'.

6.4 Conclusions

1. A systematic *in silico* analysis of human proteome identified 7 novel domains and 18 novel repeats that have not been reported earlier. Many of the domains and repeats identified were observed to be associated with disease causing proteins.

2. The 61 amino acid residue RxH domain encodes PDZ domain containing proteins which play prominent roles in synapse formation and we predict that RxH domain also has functional importance.

3. The 61 amino acid residue GLG domain is a myosin XV protein. The tails of myosin XV and myosin VIIa share several regions of amino acid identity. *Myo15* encodes an unconventional myosin (myosin XV) that is mutated in the shaker-2 (*sh2*) and shaker-2J (*sh2J*) mice, and *DFNB3*, a form of non-syndromic hearing loss in humans.

4. The 73 amino acid residue WKRK domains is associated with Williams-Beuren syndrome (WBS; OMIM 194050), that is caused by heterozygous deletions of ~1.6 Mb of chromosomal sub-band 7q11.23.

5. The 34 amino acid residue HTQ and 38 amino acid residue PTT repeats encodes the Polycystic kidney disease 1 like 3 proteins. Polycystic kidney disease (PKD) is a disease of the nephron, characterized by the formation of multiple renal tubular cysts, leading to endstage renal failure and therefore, we predict a similar function for the HTQ and PTT repeats.

6. Further database searches identified that some novel repeats and domains are also present in other mammalian genomes. Thus, the identified novel repeats and domains of human proteome can be used for annotation in the databases.

6.5 References

Alba`, M. M., Santibáñez-Koref, M. F. & Hancock, J. M. (1999a). Amino acid reiterations in yeast are over-represented in particular classes of proteins and show evidence of a slippage-like mutational process. *J. Mol. Evol.* **49,** 789–797.

Albà, M. M. & Guigó, R. (2004). Comparative analysis of amino acid repeats in rodents and humans. *Genome Res.* **14,** 549-554.

Cheney, R. E., Riley, M. A. & Mooseker, M. S. (1993). Phylogenetic analysis of the myosin superfamily. *Cell. Motil. Cytoskeleton,* **24,** 215–223.

Emili, A., Greenblatt, J. & Ingles, C. J. (1994). Species-specific interaction of the glutamine-rich activation domains of Sp1 with the TATA box-binding protein. *Mol. Cell Biol.* **14,** 1582–1593.

Francke, U. (1999). Williams-Beuren syndrome: genes and mechanisms. *Human Molecular Genetics*, **8,** 1947-1954.

Friedman, T. B., Probst, F. J., WIlcox, E. R., Hinnant, J. T., Liang, Y., Wang, A., Barber, T. D. *et al*. (2000). The myosin-15 molecular motor is necessary for hearing in humans and mice: a review of DFNB3 and shaker 2. In Berlin, C. I. & Keats, B. J. B. (eds), *Genetics and Hearing Loss*. Singular Publishing Group, San Diego, CA, pp. 31–45.

Gamsjaeger, R., Liew, C. K., Loughlin, F. E., Crossley, M. & Mackay, J. P. (2007). Sticky fingers: zinc-fingers as protein-recognition motifs. *Trends Biochem. Sci.* **32,** 63-70.

Gatchel, J. R. & Zoghbi, H. Y. (2005). Diseases of unstable repeat expansion: mechanisms and common principles. *Nat. Rev. Genet.* **6,** 743-755.

Gerber, H. P., Seipel, K., Georgiev, O., Hofferer, M., Hug, M., Rusconi, S. & Schaffner, W. (1994). Transcriptional activation modulated by homopolymeric glutamine and proline stretches. *Science,* **263,** 808–811.

Green, H. & Wang, N. (1994). Codon reiteration and the evolution of proteins. *Proc. Nat. Acad. Sci.* **91,** 4298–4302.

Hancock, J. M., Worthey, E. A. & Santibáñez-Koref, M. F. (2001). A role for selection in regulating the evolutionary emergence of disease-causing and other coding CAG repeats in humans and mice. *Mol. Biol. Evol.* **18,** 1014–1023.

Huntley, M. A. & Golding, G. B. (2002). Simple sequences are rare in the Protein Data Bank. *Proteins,* **48,** 134–140.

Imafuku, I., Waragai, M., Takeuchi, S., Kanazawa, I., Kawabata, M., Mouradian, M. M. & Okazawa, H. (1998). Polar amino acid-rich sequences bind to polyglutamine tracts. *Biochem. Biophys. Res. Commun.* **253,** 16–20.

International Human Genome Sequencing (I.H.G.S.) Consortium. (2001). Initial sequencing and analysis of the human genome. International Human Genome Sequencing Consortium. *Nature,* **409,** 860–921.

Jacobs, H. W., Knoblich, J. A. & Lehner, C. F. (1998). *Drosophila* cyclin B3 is required for female fertility and is dispensable for mitosis like cyclin B. *Genes Dev.* **12,** 3741-3751.

Jasinska, A., Wlodzimierz, J. & Krzyzosiak. (2004). Repetitive sequences that shape the human transcriptome. *FEBS Letters,* **567,** 136-141.

Karlin, S. & Burge, C. (1996). Trinucleotide repeats and long homopeptides in genes and proteins associated with nervous system disease and development. *Proc. Nat. Acad. Sci.* **93,** 1560–1565.

Karlin, S., Brocchieri, L., Bergman, A., Mrazek, J. & Gentles, A. J. (2002). Amino acid runs in eukaryotic proteomes and disease associations. *Proc. Nat. Acad. Sci. USA.* **99,** 333-338.

Kim, E. & Sheng, M. (2004). PDZ domain proteins of synapses. *Nat. Rev. Neurosci.* **5,** 771-781.

Kruglyak, S., Durrett, R. T., Schug, M. D. & Aquadro, C. F. (1998). Equilibrium distributions of microsatellite repeat length resulting from a balance between slippage events and point mutations. *Proc. Nat. Acad. Sci.* **95,** 10774–10778.

Levinson, G. & Gutman, G. A. (1987). Slipped-strand mispairing: A major mechanism for DNA sequence evolution. *Mol. Biol. Evol.* **4,** 203–221.

Li, W. H., Gu, Z., Wang, H. & Nekrutenko, A. (2001). Evolutionary analyses of the human genome. *Nature,* **409,** 847-849.

Liang, Y., Wang, A., Probst, F. J., Arhya, I. N., Barber, T. D., Chen, K., Deshmukh, D. *et al*. (1998). Genetic mapping refines *DFNB3* to 17p11.2,

suggests multiple alleles of DFNB3 and supports homology to the mouse model *shaker-2*. *Am. J. Hum. Genet*. **62,** 904–915.

Liang, Y., Wang, A., Belyantseva, I. A., Anderson, D. W., Probst, F. J., Barber, T. D., Miller, W. *et al*. (1999). Characterization of the human and mouse unconventional myosin XV genes responsible for hereditary deafness *DFNB3* and shaker-2. *Genomics,* **61,** 243–258.

Mitchell, P. J. & Tjian, R. (1989). Transcriptional regulation in mammalian cells by sequence-specific DNA binding proteins. *Science,* **245,** 371–378.

Mooseker, M. S. & Cheney, R. E. (1995). Unconventional Myosins. *A. Rev. Cell Dev. Biol.* **11,** 633–675.

Nishizawa, M., Nishizawa, K. & Kim, K. S. (1999). Tendency for local repetitiveness in amino acid usage in modern proteins. *J. Mol. Biol.* **294,** 937–953.

Nguyen, T. B., Manova, K., Capodieci, P., Lindon, C., Bottega, S., Wang, X. Y., Rogers, J. R. *et al.* (2002). Characterization and Expression of Mammalian Cyclin B3, a Prepachytene Meiotic Cyclin. *The Journal of Biological Chemistry,* **277,** 41960–41969.

Perutz, M. (1994). Polar zippers: Their role in human disease. *Protein Sci.* **3,** 1629–1637.

Ponting, C. (1997). "Evidence for PDZ domains in bacteria, yeast, and plants." *Protein Sci.* **6,** 464-468.

Probst, F. J., Fridell, R. A., Raphael, Y., Saunders, T. L., Wang, A., Liang, Y., Morell, R. J. *et al*. (1998). Correction of deafness in shaker-2 mice by an unconventional myosin in a BAC transgene. *Science,* **280,** 1444–1447.

Roussigne, M., Kossida, S., Lavigne, A. C., Clouaire, T., Ecochard, V., Glories, A., Amalric, F. *et al.* (2003). The THAP domain: a novel protein motif with similarity to the DNA-binding domain of P element transposase. *Trends in Biochemical Sciences,* **28,** 66-69.

Suzuki, K., Mori, A., Lavaroni, S., Ulianich, L., Miyagi, E., Saito, J., Nakazato, M. *et al*. (1999). Thyroglobulin regulates follicular function and heterogeneity by suppressing thyroid-specific gene expression. *Biochimie.* **81,** 329-340.

Szklarczyk, R. & Heringa, J. (2004). TRUST: Tracking Repeats Using Significance and Transitivity. *Bioinformatics,* **00,** 1-7.

Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., Smith, H. O. *et al.* (2001). The sequence of the human genome. *Science,* **291,** 1304–1351.

Wang, A., Liang,Y., Fridell, R. A., Probst, F. J., Wilcox, E. R., Touchman, J. W., Morton, C. C. *et al*. (1998). Association of unconventional myosin *MYO15* mutations with human nonsyndromic deafness *DFNB3*. *Science,* **280,** 1447–1451.

Wells, R. D. (1996). Molecular basis of genetic instability of triplet repeats. *J. Biol. Chem.* **271,** 2875-2878.

Wilkins, R. C. & Lis, J. T. (1999). DNA distortion and multimerization. Novel functions of the glutamine-rich domain of GAGA factor. *J. Mol. Biol.* **285,** 515–525.

Wilson, P. D. (2004). Polycystic kidney disease: new understanding in the pathogenesis. *The International Journal of Biochemistry & Cell Biology*, **36,** 1868-1873.

Xiao, H. & Jeang, K. T. (1998). Glutamine-rich domains activate transcription in yeast *Saccharomyces cerevisiae*. *J. Biol. Chem.* **273,** 22873–22876.

Xu, Z. X., Wang, M. R., Cai, Y., Xu, X., Han, Y. L., Wu, K. M., *et al.* (1999). Identification and expression analysis of down-regulated genes in human esophageal cancer. In *CAST (Chinese Association for Science and Technology) third academic conference of young scientists, life sciences and technology* (pp. 226–228).

Xu, Z., Wang, M. R., Xu, X., Cai, Y., Han, Y. L., Wu, K. M., *et al.* (2000). Novel human esophagus-specific gene c1orf10: cDNA cloning, gene structure, and frequent loss of expression in esophageal cancer. *Genomics*, **69,** 322–330.

Young, E. T., Sloan, J. S. & Van Riper, K. (2000). Trinucleotide repeats are clustered in regulatory genes in *Saccharomyces cerevisiae*. *Genetics,* **154,** 1053–1068.

List of publications

1) Analysis and modeling of mycolyl-transferases in the CMN group (2006). **Hemalatha Golaconda Ramulu**, Swathi Adindla and Lalitha Guruprasad. *Bioinformation.* **5:** 162-169.

2) The Rv3799-Rv3807 gene cluster in *Mycobacterium tuberculosis* genome corresponds to the 'Ancient Conserved Region' in CMN mycolyltransferases (2006). **Hemalatha G. Ramulu**, Adindla Swathi and Lalitha Guruprasad. *Evolutionary Bioinformatics Online.* **2:** 117-125.

3) Functional correlation of cylcooxygenases-1, 2 and 3 from amino acid sequences and three dimensional model structures (2006). M. Nagini, G. V. Reddy, **G. R. Hemalatha**, Lalitha Guruprasad and P. Reddanna. *Indian Journal of Chemistry*. Vol. **45A:** 182-187.

4) Identification and analysis of novel amino acid sequence repeats in *Bacillus anthracis* str. *Ames* Proteome Using Computational Tools (2007). **G. R. Hemalatha**, D. Satyanarayana Rao and Lalitha Guruprasad. *Comparative and Functional Genomics*. Volume **2007**, Article ID 47161, 23 pages.

5) Identification and Analysis of Novel Amino Acid Sequence Repeats and Domains in Pyrobaculum aerophilum Using Computational Tools (2007). **Golaconda Hemalatha**, Inampudi Krishna Kishore, Raghavarapu Srinivas Rao and Lalitha Guruprasad. *Protein & Peptide Letters*. **14:** 692-697.

6) Comparative studies of the ADAM and ADAMTS protein family members in human, frog, fly and worm genomes: A Bioinformatics Approach. Krishna Kishore Inampudi, **G. R. Hema Latha** and Lalitha Guruprasad (*being communicated*).

7) *In Silico* Identification and Analysis of Novel Amino Acid Sequence Repeats and Domains in Representative Archaeal Proteomes. **G. R. Hema Latha** and Lalitha Guruprasad (*being communicated*).

8) *In Silico* Identification and Analysis of Novel Amino Acid Sequence Repeats and Domains in Human Proteome. **G. R. Hema Latha,** Inampudi Krishna Kishore, Om Narayan, Abirami S, Shahid V. M. Prabhat Kumar and Lalitha Guruprasad (*to be communicated*).

9) Docking of small molecule inhibitors of 17 β-hydroxysteroid dehydrogenase type 10 (Human ABAD/HSD10-NAD-AG18051 complex) for elucidating the nature of interactions between ABAD/HSD10 and Aβ. **G. R. Hema Latha,** Karunakar, T. and Lalitha Guruprasad *(to be communicated).*

10) *In Silico* Identification and Analysis of Novel Amino Acid Sequence Repeats and Domains in *Azoarcus* sp. EbN1 proteome. **G. R. Hemalatha** and Lalitha Guruprasad *(to be communicated).*