STRUCTURAL AND FUNCTIONAL ASPECTS OF EVOLUTIONARILY CONSERVED SIGNATURE INDELS IN PROTEIN SEQUENCES

By

BIJENDRA KHADKA, B.Sc.

A Thesis

Submitted to the School of Graduate Studies

in Partial Fulfillment of the Requirements

for the Degree

Doctor of Philosophy

McMaster University

STRUCTURAL AND FUNCTIONAL ASPECTS OF EVOLUTIONARILY CONSERVED SIGNATURE INDELS IN PROTEIN SEQUENCES

DOCTOR OF PHILOSOPHY (2019) McMaster University

(Biochemistry) Hamilton, Ontario

TITLE: Structural and Functional Aspects of Evolutionarily Conserved Signature Indels in Protein Sequences.

AUTHOR: Bijendra Khadka, B.Sc. (McMaster University)

SUPERVISOR: Professor Dr. Radhey S. Gupta

NUMBER OF PAGES: xxii, 189

DEDICATION

"Lovingly dedicated to my Mother"

ABSTRACT

Analysis of genome sequences is enabling identification of numerous novel characteristics that provide valuable means for genetic and biochemical studies. Of these characteristics, Conserved Signature Indels (CSIs) in proteins which are specific for a given group of organisms have proven particularly useful for evolutionary and biochemical studies. My research work focused on using comparative genomics techniques to identify a large number of CSIs which are distinctive characteristics of fungi and other important groups of organisms. These CSIs were utilized to understand the evolutionary relationships among different proteins (species), and also regarding their structural features and functional significance. Based on multiple CSIs that I have identified for the PIP4K/PIP5K family of proteins, different isozymes of these proteins and also their subfamilies can now be reliably distinguished in molecular terms. Further, the species distribution of CSIs in the PIP4K/PIP5K proteins and phylogenetic analyses of these protein sequences, my work provides important insights into the evolutionary history of this protein family. The functional significance of one of the CSI in the PIP5K proteins, specific for the *Saccharomycetaceae* family of fungi, was also investigated. The results from structural analysis and molecular dynamics (MD) simulation studies show that this 8 aa CSI plays an important role in facilitating the binding of fungal PIP5K protein to the membrane surface. In other work, we identified multiple highly-specific CSIs in the phosphoketolase (PK) proteins, which clearly distinguish the bifunctional form of PK found in bifidobacteria from its homologs (monofunctional) found in other organisms. Structural analyses and docking studies with these proteins indicate that the

CSIs in bifidobacterial PK, which are located on the subunit interface, play a role in the formation/stabilization of the protein dimer. We have also identified 2 large CSIs in SecA proteins that are uniquely found in thermophilic species from two different phyla of bacteria. Detailed bioinformatics analyses on one of these CSIs show that a number of residues from this CSI, through their interaction with a conserved network of water molecules, play a role in stabilizing the binding of ADP/ATP to the SecA protein at high temperature. My work also involved developing an integrated software pipeline for homology modeling of proteins and analyzing the location of CSIs in protein structures. Overall, my thesis work establishes the usefulness of CSIs in protein sequences as valuable means for genetic, biochemical, structural and evolutionary studies.

ACKNOWLEDGMENTS

First and foremost, I would like to express my profound gratitude to my supervisor, Professor Dr. Radhey. S. Gupta, whose mentorship, patient guidance and steadfast support have helped shape my thesis. His dedication and contribution to the field of science are immense, and I have always inspired by his elegant and interesting research ideas. The door to his office was always open to me whenever I needed his scholarly advice and motivation at difficult times. I greatly appreciate the independence he gave me to learn several techniques and to spearhead different projects which made my Ph.D. experience productive and stimulating. His encouragement and attention to detail have significantly improved my research over the years and I am very grateful to him for the generous opportunity to be a part of his research lab and chance to work with him and to learn from him which allowed me to grow as a research scientist.

I would also like to extend my warmest gratitude to my Supervisory Committee members Professor Dr. Herb E. Schellhorn and Professor Dr. Richard Epand for their time and for continuously providing me with invaluable scientific advice, wisdom and insightful discussions on the various aspects of my thesis projects. I also wish to express my gratitude to my previous committee member Professor Dr. Joaquin Ortega, for his guidance and support.

I also must thank and acknowledge the contributions of all the past and present members of Dr. Gupta's Laboratory: Dr. Sohail Naushad, Dr. Mobolaji Adeolu, Nazmul Hassan and Bashudev Rudra who made my work and stay at the lab immensely pleasurable. I am particularly grateful to Sohail Naushad and Mobolaji Adeolu, for their

intellectual support throughout the course of my Ph.D. program. I greatly enjoyed the good-spirited discussions with them related to the research.

I thankfully acknowledge current and past undergraduate research students from Dr. Gupta's lab. In particular, Seema Alnajar, Anish Nanda, Dhillon Perusad, Sadisha Galappatti, and Ryan Martin for their collaboration and contribution to the various aspects of the projects related to my thesis.

I would like to show my appreciation to all the faculty members and staff of the Department of Biochemistry and Biomedical Sciences at McMaster University who accompanied me throughout these years. I gratefully acknowledge Lisa Kush and Tylor Allison who graciously provided me with the administrative and technical assistance.

Outside the lab, I have been very fortunate to have found many friends in Hamilton and I am thankful for all the good times that we spent together and for making my stay a memorable experience. I would like to acknowledge my friends Phyllis Cameron Ung, Matt Mclean (Skelly) and Paul Kolb for their wholehearted and unwavering support. Thanks to Uday Shah and Sagar Bhatta (you guys have always been nice to me). Thanks to Sanjna Pradhan for everything.

Finally, I would like to thank my parents for believing in me and for selflessly providing me with the opportunity and freedom to succeed, and the encouragement to explore, follow and pursue my dream. Especially my mother for always having been supportive of me despite the distance and for her unconditional love and patience.

PREFACE

The following work is a sandwich thesis. Chapter 1 provides an introduction to the different subjects to provide contexts for the significance of the manuscripts and Chapters described in this thesis. Chapters 2, 3, and 4 are unaltered manuscripts published in the years 2017 to 2019. Chapter 5 is an unaltered manuscript submitted for publication in November 2019. Chapter 6 describes an in-house developed software pipeline for homology modelling of CSI-containing proteins. Chapter 7 reflects on the presented studies and describes the overall usefulness and future directions of the work. References for Chapters 1, 6 and 7 are provided at the end of this thesis. The preface section in each Chapter describes the details of the published and submitted work, as well as my contribution to the multiple-authored articles. All the chapters have been reproduced with the consent of all co-authors. Irrevocable, non-exclusive license has been granted to McMaster University and to the National Library of Canada from all publishers. Copies of permission and licenses have been submitted to the School of Graduate Studies.

TABLE OF CONTENTS

LIST OF FIGURES

CHAPTER 2

CHAPTER 3

CHAPTER 4

CHAPTER 5

CHAPTER 6

LIST OF TABLES

CHAPTER 2

CHAPTER 4

CHAPTER 6

LIST OF ABBREVIATIONS

2D ..... Two-Dimensional
3D ..... Three-Dimensional
aa ..... amino acids
BLAST ..... Basic Local Alignment Search Tool
BLASTp ..... Protein vs Protein BLAST search
CSI ..... Conserved Signature Indel
CSP ..... Conserved Signature Protein
DGK ..... Diacylglycerol Kinase
DNA ..... Deoxyribonucleic acid
DnaK ..... Chaperone DnaK (Hsp60)
DOPE ..... Discrete Optimized Potential Energy
DPPC ..... Dipalmitoyl-Phosphatidylcholine
EPS ..... Electrostatic Potential Surface
F6P Fructose-6-phosphate
G6P Glucose-6-phosphate
GlabModeller ..... Gupta Lab Modeller
GI ..... GeneBank Identifier
GLEANS ..... Gupta Lab Evolutionary Analysis Software
GLU ..... Glutamic acid
GroEL ..... Chaperonin GroEL (Hsp70)
GUI ..... Graphical User Interference
HGT ..... Horizontal Gene Transfer
Hsp60 ..... Heat shock protein of the 60 kDa size
Hsp70 ..... Heat shock protein of the 70 kDa size
Indel ..... Insertion/Deletion
MAFTT ..... Multiple Alignment using Fast Fourier Transform
MD ..... Molecular Dynamics
MEGA ..... Molecular Evolutionary Genetics Analysis
ML ..... Maximum-Likelihood
MSA ..... Multiple Sequence Alignment
MUSCLE ..... MUltiple Sequence Comparison by Log-Expectation
NCBI ..... National Center for Biotechnology Information
NJ ..... Neighbour-Joining
PDB ..... Protein DataBank
PI4P Phosphatidylinositol-4-Phosphate
PI5P Phosphatidylinositol-5-Phosphate

PIP4K Phosphatidylinositol-5-phosphate-4-kinase
PIP5K Phosphatidylinositol-4-phosphate-5-kinase
PIPER ..... Fast Fourier Transform (FFT)-based protein docking program
PK ..... Phosphoketolase
PtdIns ..... Phosphatidylinositols
PI(4)P Phosphatidylinositol-4-Phosphate
$PI(4,5)P_2$ Phosphatidylinositol-4,5-Bisphosphate
POPC ..... Palmitoyl-oleoyl Phosphatidylcholine
PSI-BLAST ..... Position-Specific Iterated BLAST
PyMOL ..... Python-enhanced Molecular Graphics Tool
RC ..... Reaction Center
RMSD ..... Root-mean-square-deviation
RMSF ..... Root-mean-square-fluctuation
RNA ..... Ribonucleic acid
RpoA ..... DNA dependent RNA polymerase α-subunit
RpoB ..... DNA dependent RNA polymerase β-subunit
SecA ..... Secretory A, a conserved ATPase protein
SSU rRNA ..... Small subunit ribosomal ribonucleic acid
Tk ..... Tool Kit
UMCA ..... Unicellular Metazoan Common Ancestor
X5P Xylulose-5-phosphate
XFPK ..... X5P/F6P Phosphoketolase only found in Bifidobacteria
XPK ..... X5P Phosphoketolase found in other bacteria
Å ..... Angstrom, 0.1nm

GLOSSARY

Ancestor: Any organism, population or species from which some other organism, population or species is descended.

Apomorphy: Specialized or derived character-state of an organism.

Bifid Shunt or Fructose-6-phosphate pathway: A central carbohydrate catabolic pathway unique to Bifidobacteria and which relies on an enzyme phosphoketolase to catalyze fructose-6-phosphate (F6P).

Bilayer: A back-to-back arrangement of monolayers of lipid molecules with non-polar hydrophobic tails of the lipids faces inwards and their hydrophilic polar head groups arrayed on the bilayer surface.

Bootstrapping: A statistical procedure to assess the reliability of a result (usually a phylogenetic tree) that involves sampling data into a given number with replacement form the original set.

Choanoflagellates: Unicellular protests phylogenetically closest to the metazoans.

Clade: A monophyletic group composed of an ancestor and all of its descendants.

Comparative Genomics: A field of biological research that compares genomic features of various organisms/species such as sequence characteristics, genes, proteins, gene order, regulatory sequences, and other genetic or molecular characteristics in order to reveal the biological and evolutionary relationships and differences between organisms.

Conserved Signature Indel (CSI): Insertion or deletion of a specific size uniquely present in a specific region in gene/protein sequences of organisms from the group of interest and absent in all other bacterial groups. Conserved residues flanked on both sides ensure its reliability.

Convergent Evolution: The independent evolution of similar traits in distantly related organisms due to adaptive benefits to similar environments.

Duplication: Mechanism through which sequence is duplicated during molecular evolution.

Eukaryote: One of the three domains of life, differentiated from prokaryotes by the presence of a nucleus or other membrane-bound organelles.

Fungi: A group of saprophytic and parasitic spore-producing eukaryotic organisms that are grouped in a distinct kingdom within the eukaryotes.

GUI: Graphical User Interface (pronounced Gooey), is a visual component of software that relies on pictures, windows, icons, and menus to direct the interaction of users with applications.

Homologs or Homologous genes/proteins: Sequences that are evolutionarily related by descent from a common ancestor.

Homology Model: The three-dimensional (3D) structure of a protein (query) generated from its amino acid sequence and experimental 3D structures of evolutionarily related proteins that share a similar structure.

Horizontal Gene Transfer: Transfer of genetic materials between organisms other than by descent in which transmission of DNA occurs through the generations as the cell divides.

Isoforms or Genes/Proteins Isoforms: Sequences that are similar to each other and that have arisen from the same sequence or different sequence as a result of alternative splicing.

Metazoa: Multicellular eukaryotic organisms with differentiated cells and tissues.

Maximum likelihood tree: A phylogenetic tree built using the maximum likelihood method that searches for the tree topology that has the maximum probability of being produced by the given alignment.

Molecular Dynamics: A computational approach that allows to study the time evolution of a system of interacting particles (atoms, molecules, etc.).

Multiple Sequence Alignments (MSA): Representation of two or more sequences in such a way that reflects their relationships.

Orthologs or Orthologous genes/proteins: Sequences from different species that are evolutionarily related by descent from a common ancestral sequence and that diverged from one another as a result of speciation divergent events.

Paralogs or Paralogous genes/proteins: Sequences within the same organism/species and that result from duplication of one original sequence.

Phylogenetic tree: Representation of evolutionary relationships between a set of sequences, species or organisms, etc.

Phosphoinositides: A family of minority acidic phospholipids located in the cytosolic face of the eukaryotic cell membranes.

Protein Data Bank: The repository of experimentally determined three-dimensional structural data of large biological molecules, such as proteins, nucleic acids, and complex biomolecular assemblies.

Protein Family: A group of proteins that share a common evolutionary origin, reflected by their related functions and similarity in sequences or structures.

Ramachandran Plot: A scatterplot depicting the disposition of backbone phi (φ) and psi(ψ) torsion angles for each residue in a protein or set of proteins. It is a fundamental tool in structural biology for the analysis of protein structures.

Root Mean Square Deviation (RMSD): The measure of the average distance between the atoms (usually the backbone or Cα atoms of the entire protein) of superimposed proteins or residues. It indicates the overall flexibility of protein during MD simulation.

Root Mean Square Fluctuation (RMSF): The measure of the fluctuation of the atoms (usually the backbone or Cα atoms of the individual residues of protein) coordinates from their average position. It indicates the structural flexibility of each amino acid in a protein during MD simulation.

Single-gene/protein phylogenetic tree: Reconstruction of a phylogenetic tree based on the comparison of homologous sequences representing a single gene or protein.

Synapomorphy: A derived character-state, and because it is shared by the taxa under consideration, is used to infer common ancestry.

Unicellular Metazoan Common Ancestor: The single-celled ancestors of metazoan from which all the metazoans are derived.

CHAPTER 1

BACKGROUND AND INTRODUCTION

1. From Organisms to Molecular Sequences to understand Evolutionary Relationships

Understanding the origin and evolutionary relationships among organisms constitutes a formidable challenge in biological sciences. The elements of evolutionary thought can be traced back to Greek philosophers like Aristotle (384-322BC) who in his *scala nature* ("ladder of life") classified organisms hierarchically based on the observed common attributes (e.g. blooded and bloodless), with inanimate things through the plants to the bottom and higher animals up to man at the pinnacle of creation (Mayr, 1982; Kullmann, 1991; Ragan, 2009). The modern basis for the ranked-based classification of living systems was first purposed by Carl Linnaeus in the 18th century, in his book "*Naturae Sytemae"* (Linneaus, 1758). However, the first systematic studies for the understanding of the evolutionary relationship between organisms must be assigned to the seminal works of Charles Darwin who provided insights into how the evolutionary process works to generate different life forms (Darwin, 1859). Several great steps were made throughout the second half of the 18th century and continuing through to the first half of the 20th century that led to the methods based solely on morphological, physiological and biochemical characteristics for analyzing the relationship within and among organisms (Cohn, 1875; Vaughan, 1906; Stanier and Van Niel, 1941; Sapp, 2009; Oren, 2010; Oren and Garrity, 2014; Ramasamy et al., 2014). These long-established methods used for the biological classification, however, began to approach their explanatory limits in the second half of the 20th century due to the plastic, analogous, and often convergent nature of the examinable characteristics available to the scientists at the

time (Stanier et al., 1963; Woese, 1987; Gupta, 1998; Oren, 2010). The advent of the ability to determine nucleic acid, molecular sequences and structural data by the development of powerful experimental, computational and mathematical methods offered a novel approach to infer relationships and evolutionary history of genes and organisms (Watson and Crick, 1953; Crick, 1958; Sanger, 1959; Zuckerkandl and Pauling, 1965; Eck and Dayhoff, 1966). Since then the use of molecular data to understand the evolutionary relationships among organisms has proven to be a more consistent and objective approach to classification than morphological and biochemical approaches. By early 1960, the key practices of molecular evolution, which included collecting, comparing, and computing sequences were already developed (Sanger, 1949; Brown et al., 1955; Harris et al., 1956; Crick, 1958; Margoliash et al., 1959; Sanger and Tuppy, 1951). The notion of comparing the molecular sequence to infer relationship was later strengthened by the work of Zuckerkandl and Pauling who put forward the compelling idea of using the molecular sequences as a document of the evolutionary history of an organism and to deduce phylogenetic relationships (Zuckerkandl and Pauling, 1965). This marked the beginning of the field of molecular evolution, which later acquired enormous momentum with a major improvement in the techniques, such as nucleotide sequencing, utilized to characterize the molecular basis of genetic changes (Edman and Begg, 1967; Sanger et al., 1977). The astonishing power of molecular data didn't come to light until Woose and colleagues revelation, based on the use of sequences from an SSU rRNA (also known as 16S or 18S rRNA to denote size), of the presence of yet another domain of life besides bacteria and eukaryotes, which they called as *the Archaebacteria* (later termed as

the Archaea) (Woese and Fox, 1977; Woese et al., 1990). The significant progress in sequencing which brought the dawn of molecular data continues to provide the potential to resolve important evolutionary relationships among organisms and proteins.

2. Impact of Molecular Sequence and Structures on Comparative Evolutionary Studies

Even before the first protein structure was resolved, it was realized that the primary sequence of proteins carry information, and it can form structurally discrete and ordered motifs (secondary structure elements), which would then arrange compactly into the more functional three dimensional structural form (Pauling et al., 1951; Lindorff-Larsen et al., 2012; Bragg et al., 1950). The connecting link between the sequence and structural relationship was hotly contested for decades and was later established by Christian Anfinsen and colleagues (Anfinsen et al., 1961; Anfinsen, 1973). Soon after the first structure of a protein was reported, Anfinsen in his book "The Molecular Basis of Evolution" wrote: "*A comparison of the structures of homologous proteins (i.e., proteins with the same kinds of biological activity or function) from different species is important, therefore, for two reasons. First, the similarities found to give a measure of the minimum structure which is essential for biological function. Second, the differences found may give us important clues to the rate at which successful mutations have occurred throughout evolutionary time and may also serve as an additional basis for establishing phylogenetic relationships.*" (Anfinsen, 1959) pg. 143). Through their experiment on enzyme ribonuclease, they elegantly demonstrated that the sequence information of a

protein contains all the information sufficient for the folding of a protein into its native spatial structure (Anfinsen and Haber, 1961; Anfinsen et al., 1961; Anfinsen, 1973). The major breakthrough in understanding the relationship between sequence and structure of proteins awaited the discovery of X-crystallography and its successful application to nucleic acid and proteins (Bragg and Bragg, 1913; Campbell, 2002; Watson and Crick, 1953; Bernal and Crowfoot, 1934). The first crude model of a protein 3D structure of myoglobin was reported by Kendrew in 1958 (Kendrew et al., 1958). Subsequently, in 1960, the 3D structure of haemoglobin in atomic detail was solved (Perutz et al., 1960; Perutz, 1985). The success of myoglobin and haemoglobin structures revolutionized the field of structural biology and furthered our mechanistic understanding of biology. In the decades that followed, a barrage of structural information was obtained for several additional protein molecules, most notably including the atomic structure of a first enzyme - hen white lysozyme (Blake et al., 1965), ribonuclease (Kartha et al., 1967; Avey et al., 1967), carboxypeptidase A (Lipscomb et al., 1969), etc. These early structures, together with those of many other enzymes in the 1970s and beyond, revealed the mechanistic details about the protein function such as active site conformations and catalytic mechanisms. At the same time, the influx of structural data (e.g. X-ray crystallography) also created demand for computerized methods for the systematic examination of the complex molecules and laid the foundation of the field of Structural Bioinformatics. The structural information available at the time allowed comparative analysis to be carried out on related proteins and provided insight into the subsequent effects of change in amino acid sequences such as insertions and deletions in three-

dimensional structure and function (Zuckerkandl and Pauling, 1965; Almassy and Dickerson, 1978; Lesk and Chothia, 1980; Chothia and Lesk, 1986). These earlier studies provided the basis for the understanding of the protein sequence, structure and functional relationship and fueled the beginning of comparative protein structure modelling approaches (Levinthal, 1966; Browne et al., 1969; Hartley, 1970; Greer, 1981; Perutz, 1983; Chothia and Lesk, 1986) which would later mature into the fully automated pipelines with the advent of computational methods (Sali et al., 1990; Fiser, 2004). Later, the culmination of decades of biochemical and structural studies and growing number of structural data led to the foundation of the Protein Data Bank (PDB) as a repository for the deposition of three-dimensional (3D) coordinates of experimentally determined biological macromolecules (Protein Data Bank, 1971). Established in 1971 with just seven inaugural structures, the PDB now holds the information for 154,015 structures (as of July 2019) with the number growing rapidly every year (Bernstein et al., 1977; Berman et al., 2003; Burley et al., 2019). More recently, in addition to the wealth of structural and sequence data, we have witnessed the development of sensitive tools for sequence and structural similarity searches that are beginning to shed light on the evolution of protein structures and functions.

3. The rise of the Genomics Era and its Implication on Comparative Genomics and Evolutionary Systematics

The first revolution in DNA sequencing took place nearly four decades ago (Heather and Chain, 2016; Shendure et al., 2017). Two methods, the Maxam-Gilbart

chemical cleavage, and the Sanger-Coulson chain termination marked an important early landmark that would forever alter the development of sequencing technology (Maxam and Gilbert, 1977; Sanger et al., 1977; Gilbert and Maxam, 1973). Despite the extremely complex and laborious nature of these early techniques, they greatly enhanced the throughput sequencing of DNA. In the following several decades, these methods were commercialized and scaled up through the concerted efforts of several research groups bringing them to dominate the field of DNA sequencing. This in turn allowed for the sequencing of the several landmark genomes (Staden, 1979; Messing et al., 1981; Sanger et al., 1982; Smith et al., 1986; Connell et al., 1987). This was followed by exponential growth in sequence data from various genes/proteins and organisms, and motivated researchers for the creation of central open access data repositories, such as GenBank (Burks et al., 1985; Kneale and Bishop, 1985; Bilofsky and Burks, 1988) and European Molecular Biology Laboratory (EMBL) data library (Kneale and Kennard, 1984; Cameron, 1988). The later development and implementation of various search algorithms such as BLAST (Altschul et al., 1990; Butler, 1993), greatly enriched the spirit of data sharing and the value of each deposited sequences. By 1995, the first complete genome of *Haemophilus influenzae* was available, which marked the beginning of a genomic era (Fleischmann et al., 1995). As of present writing of this thesis, almost 25 years after the sequencing of *H. influenzae*, over 100,000 complete and draft bacterial genomes can be found in the GenBank database. Sequencing of the draft human genome was not completed until the turn of the millennium and ultimately cost over 2.7 billion USD and took over decades to complete (Yamey, 2000; Hood and Rowen, 2013; Emmert-Streib et

al., 2017). This development marked the arrival of modern sequencing techniques, the so called "*next generation"* sequencing (NGS) or "second-generation (2G)" methods such as the Illumina DNA sequencing platforms, HiSeq and MiSeq (van Dijk et al., 2014; Caporaso et al., 2012). In the following years, these breakthrough platforms were rapidly joined by complimentary third-generation (3G) methods, such as Pacific Biosciences (PacBio) single-molecule real-time (SMRT) sequencing approach (Eid et al., 2009; Schadt et al., 2010; van Dijk et al., 2018) and to fourth-generation (4G) methods, such as Oxford Nanopore Technologies nanopore-based single molecule sequencing technology (Ke et al., 2016; Jain et al., 2018; van Dijk et al., 2018). The development of more efficient screening platforms and the growing competition among several vendors to develop faster and more cost-effective machines has led to a steady decline in sequencing cost and increase in sequencing speed by several orders of magnitude. This ultimately contributed to the fulfillment of the original goal of sequencing the genome for less than $1000 USD (Schloss, 2008; van Dijk et al., 2014). Recent progress in efficient sequencing methods has led to a remarkable feat of sequencing the whole human genome in 19 hours (Clark et al., 2019). The rapid reduction in sequencing cost and time has led to its democratization, bringing the genome sequencing technology within reach of small laboratories and researchers around the globe and resulting in the generation of complete genome sequences from a wide range of biologically and medically important organisms (Shendure and Ji, 2008; Kyrpides, 2009; Parkhill and Wren, 2011; Shendure et al., 2017). With well over 100,000 bacterial genomes and hundreds of eukaryotic genomes sequences currently available on public repositories and many thousands more from

diverse organisms in the pipeline, the information for genome sequences will continue to grow in years to come.

This growing wealth of genomic data has allowed for the development of several novel and powerful analysis methods with a multitude of potential current and future applications. It has also provided a new dimension for understanding the evolutionary relationship between organisms (Gupta, 1998; Koonin et al., 2000; Charlesworth et al., 2001; Wei et al., 2002; Chun and Rainey, 2014; Tyzack et al., 2017). One such method of analysis is comparative genomics, which incorporates both the creation of computational tools and using those tools to delimit the genetic basis of diversity of organisms and strains (Wei et al., 2002). Another important aspect of comparative genomics is that it provides insights on the pathogenesis of organisms, and also offers vast potential toward identification of novel drug targets for the development of novel antimicrobial agents (Moir et al., 1999; Loferer, 2000; Cole, 2002; Kramer and Cohen, 2004; Klemm and Dougan, 2016; Sharma et al., 2019).

Among the most widely used comparative genomics methods include examining the overall nucleotide statistics (e.g. overall (G+C) content) (Alm et al., 1999), the analysis of syntentic relationships, the comparison of gene locations, relative gene order and regulation (Snel et al., 1999; Belda et al., 2005), the comparison of protein content (Pellegrini et al., 1999; Tatusov et al., 1997), core and pan-genome analysis (Tettelin et al., 2005; Medini et al., 2005; Tettelin et al., 2008), the identification of genome signatures (Campbell et al., 1999; Gusev et al., 2014) and the construction of supertree and super-matrix based phylogenetic trees (Sanderson et al., 1998; de Queiroz and

Gatesy, 2007). Comparative genomics also involves analysis of events such as gene loss, gene duplication and horizontal gene transfer (HGT) (Tatusov et al., 1997; Li et al., 2019). These types of analyses are aimed at extending beyond a mere description of differences and similarities between organisms and are focused toward developing the models and theories that could explain such phenomenon. The growing availability of multiple genome sequences from diverse model and non-model organisms has allowed comparative genomic analysis to be carried out at increasingly larger scales in recent years, shedding light on various functional and evolutionary questions that were previously out of reach. These approaches have proven useful for inferring complex evolutionary history and phylogenetic relationships among organisms, which is central to the advancement of our understanding of organism diversity and evolution. Although, the available comparative methods are useful for inferring relationships among organisms, however, they are all based on the principle of measuring the degree of relatedness or similarity of genomes, rather than providing unique distinguishing characteristics features that may differentiate a group of related organisms. In this light, there is a need for the identification of novel specific molecular, biochemical and genetic characteristics derived from the growing genome sequences which can serve as distinctive molecular markers for a robust interpretation of the relationship between a related group of organisms (Sutcliffe et al., 1992; Gupta, 1998; Gupta and Griffiths, 2002; Klenk and Goker, 2010; Verma et al., 2013; Whitman, 2015; Gupta, 2016a; Chun et al., 2018; Sutcliffe, 2015).

4. Evolutionary Studies Using Molecular Sequences and Phylogenetics

Several decades before DNA sequencing became feasible, it was realized that analysis of amino acid or nucleic acid sequences data could be used to decipher the evolutionary history of molecules (Eck, 1962; Zuckerkandl and Pauling, 1965; Eck and Dayhoff, 1966). The pioneering work of the late Margaret Dayhoff and other researchers during the 1960s had already enabled an early form of bioinformatics, allowing researchers to utilize computation in sequence determination and comparing sequences from multiple organisms (Dayhoff, 1965; Fitch and Margoliash, 1967; Needleman and Blair, 1969; Doolittle and Blombaeck, 1964). Since then, the application of bioinformatics approaches has been increasingly involved in studying protein structure, function, and evolution (Hagen, 2000). These bioinformatics approaches include methods, tools, and algorithms for efficient sequence alignment, database searches to identify homologous sequences and structures, and to the generation of phylogenetic trees to decipher evolutionary relationships. In recent years, unparalleled advancements in computing methods and algorithms in combination with cheap sequencing costs have led to the accumulation of an abundance of genomics data. This wealth of information has further enhanced our ability to carry out more powerful comparative analysis to address various key biological and evolution related research questions (Tatusov et al., 1997; Wei et al., 2002).

Modern phylogenetic analyses most often relied on the usage of both nucleotide and protein sequences. However, protein-based phylogenetic analyses are thought to be more reliable than nucleotide-based analyses. This is due to the fact that some

phylogenetic trees based on nucleotide sequences are suggested to be misleading due to factors such as the difference in G+C content among lineages and the effect of the degeneracy of genetic codes on nucleotide sequences (Karlin et al., 1995; Gupta and Johari, 1998). Phylogenetic analysis using protein sequences usually starts with the identification of homologs/orthologs from related family members in the protein databases. Once the sequences are retrieved, they are aligned, and the multiple sequence alignment (MSA) obtained forms an essential preliminary for phylogeny reconstruction. The alignment step constitutes a very crucial step for evolutionary study of proteins, as the improvement in MSA have been shown to improve the phylogenetic accuracy, although these improvements are minor (Cantarel et al., 2006; Hall, 2005; Ogden and Rosenberg, 2006). Despite these challenges, several MSA algorithms exist with improved speed and reasonable accuracy such as Clustal series of programs (e.g. ClustalX and Clustal Omega) (Sievers et al., 2011; Chenna et al., 2003), MUSCLE (Edgar, 2004), MAFTT (Katoh et al., 2005) and T-Coffee (Notredame et al., 2000). For phylogenetic reconstructions, the most commonly applied methods include distance-matrix methods and character-based methods. Distance-matrix methods compute a matrix of pair-wise genetic “distance” between sequences and summarize it using the hierarchical clustering algorithm such as Unweighted Pair-Group Method with Arithmetic Mean (UPGMA) (Sneath and Sokal, 1973) or Neighbor-Joining (NJ) (Saitou and Nei, 1987). Compared to other methods, NJ or UPGMA have the advantage of being rapid and are therefore the method of choice when carrying out large-scale phylogenetic analyses. In contrast, character-based methods attempt to infer the phylogeny based on all the individual

characters in a simultaneously compared alignment of sequences. These methods include maximum parsimony (MP), which strives for the tree with an overall minimum number of overall genetic changes between the taxa by randomly changing the topology of the tree until the parsimony is no longer improved (Fitch, 1971), the Maximum-likelihood (ML), strives for the tree with the maximal likelihood to produce the variation observed in the given set of sequence data (Felsenstein, 1981), and the Bayesian inference method, which seeks to obtain the full posterior probability distribution of all possible phylogenies by combining the prior probability distribution with the tree likelihood of evolutionary parameters (Huelsenbeck et al., 2001). When considering the inferring process of sequence evolution, the ML method possesses a clear advantage over a distance or parsimony methods, as it utilizes more of the information content of the underlying sequences, such as positional variability, transitions/transversion ratio, character state probability per position and many others (Felsenstein, 1981). The major drawback of the ML method, however, is that it is computationally demanding. Since the inference from the individual phylogenetic outcomes from most of the aforementioned methods has been suggested to be affected by several factors, the most common practice is to include statistical tests such as bootstrapping as a part of a tree evaluation method for a through phylogenetic analysis (Felsenstein, 1985; Efron, 1992; Tateno et al., 1994). Bootstrapping is a statistical approach to measure the robustness of a phylogenetic tree that involves random sampling of data into a given number of bootstrap sets with replacement from the original data. One major advantage of this approach is that it can be applied to distance, parsimony, likelihood and just about any of the other tree-constructions methods,

although it must be noted that applying this method can be computationally intensive due to the increase in the number of bootstrap samples requested.

A phylogenetic tree based upon a single gene or its protein sequence only reflects the evolution of that particular gene, not the evolutionary history of the organisms from which the gene or proteins was isolated. The true phylogenetic history of organisms can be more accurately elucidated using the sequence data from multiple genes as well as using the biochemical, morphological and physiological characteristics of organisms. Overall, the use of phylogenetic analyses using molecular sequences has been central for several evolutionary studies. In addition to the conventional taxonomic use to infer the evolutionary relationships between organisms, phylogenetic trees are often applied, among many others, to detect horizontal gene transfers (HGT) (Bielawski and Yang, 2004; Ravenhall et al., 2015), to infer orthology and paralogy relationships between proteins (Fitch, 2000; Gabaldon and Koonin, 2013), to investigate gene duplication events (Donoghue and Mathews, 1998; Philippon et al., 2015) or to infer the ancestral state of genes/proteins (Chang et al., 2005; Hall, 2006).

5. Conserved Signature Indels as a Tool for Evolutionary Studies

The advent of DNA sequencing technology, coupled with advanced computational methods in the past two decades has led to the acquisition of a significant abundance of genome sequence information from diverse organisms (McPherson, 2014). The wealth of genomic information provides an unparalleled opportunity to carry out various studies to discover novel molecular characteristics specific for a related group of organisms. These

shared molecular signatures that are ideal for evolutionary studies should be homologous apomorphic characters introduced only once during the course of evolution (Stackebrandt and Schumann, 2006; Gupta, 2014). One class of such molecular markers that our lab has pioneered the discovery and usage of, and has been a focus of recent evolutionary studies, is Conserved Signature Insertions and deletions (viz. Indels) (CSIs) that are found in a conserved region of the protein homologs from a particular group of organisms (Rivera and Lake, 1992; Baldauf and Palmer, 1993; Gupta, 1998; Rokas and Holland, 2000; Gupta and Griffiths, 2002; Gupta, 2014; Naushad et al., 2015; Gupta, 2016b; Zhang et al., 2016a). CSIs that serve as a useful molecular markers are the regions in sequence alignments where a specific change is observed in the primary structure of particular proteins in all members of one or more defined groups of species but not in other groups (Gupta, 1998). The signatures must be flanked on both sides by conserved regions to ensure their reliability and to rule out alignment artifacts or errors (Gupta, 2014). Since these CSIs are limited to a specific group of organisms, the simplest and most parsimonious explanation for the presence of these CSIs in all member of a particular group is that the rare genetic change that gave rise to the CSIs occurred once in a common ancestor and it was then vertically inherited by various descendants (Rivera and Lake, 1992; Baldauf and Palmer, 1993; Gupta, 1998; Rokas and Holland, 2000; Gupta and Griffiths, 2002; Gupta, 2014) Additionally, based on the presence or absence of these signatures in different out-group species, it is also possible to infer whether a given CSI is an insertion or a deletion (Gupta, 1998; Gupta, 2014). However, it is also important to consider the possibility that, in some cases, the shared presence of these CSI

could also result from non-specific mechanisms such as lateral gene transfer (LGTs) or from independent occurrence of similar genetic change in evolutionary unrelated lineages (convergent evolution) (Gupta et al., 2017; Gupta, 2018).

A well-defined CSI in a particular protein also serves as an important milestone for evolutionary events, since all the descendent containing this protein are expected to share the CSI whereas the homologous protein in all other organisms which existed before this event will lack the CSI (Rivera and Lake, 1992; Baldauf and Palmer, 1993; Gupta, 1998; Rokas and Holland, 2000; Gupta and Epand, 2017). Further, the presence or absence of CSIs in various lineages or proteins is generally not affected by factors such as differences in evolutionary rates among lineages or proteins, variable artifacts affecting the construction of phylogenetic trees (Gupta, 1998; Rokas and Holland, 2000; Gupta, 2016b). Therefore, even a CSI of 1 aa length provides a very useful and reliable marker for evolutionary studies (Gupta, 1998; Singh and Gupta, 2009). In some cases, when two proteins are evolved as a result of ancient gene duplication, the presence or absence of the CSI in the homologous protein has also proven useful to ensure whether the observed CSI is an insertion or a deletion (Gupta, 1998; Valas and Bourne, 2009). Over the last couple of decades, Dr. R. S. Gupta and his group has utilized these CSIs to resolve important aspect of the microbial phylogeny and systematics (Gupta, 1998; Rokas and Holland, 2000; Gupta, 2014; Gupta, 2016b; Gupta et al., 2017; Khadka et al., 2017; Khadka and Gupta, 2019). The utility of CSI to shed insight into a number of important evolutionary questions is discussed in Chapters 2, 3, 4 and 5 of this thesis.

6. Importance of Structural and Functional Studies on CSIs in Protein Structure

Earlier work from our lab on a number of CSIs in essential proteins (e.g., Hsp70 and Hsp60) has provided substantial evidence that they play a critical for cell growth in the organisms for which they are found (Singh and Gupta, 2009). However, due to a lack of structural information for most of the proteins in which CSIs are found, the structural and functional characteristics of these CSIs remains largely unknown. Analysis of available three-dimensional (3D) structures of proteins show that most of the identified CSIs are generally found on the surface loops of proteins and that they are usually located away from the active site which suggests that they do not disrupt the core function of a protein (Singh and Gupta, 2009; Gupta, 2010; Gao and Gupta, 2012; Gupta, 2014; Gupta and Khadka, 2015; Gupta, 2016b; Gupta et al., 2017; Khadka and Gupta, 2017; Alnajar et al., 2017). The occurrence of CSIs on surface loops also suggests that they could act both as an "enabling loops" or a "disabling loops" (Akiva et al., 2008). The enabling features of indels could be useful to mediate specific interaction (e.g., dimer stabilization, or binding to some other proteins or ligands), whereas the disabling features of indels could prevent interactions with unwanted partners (Akiva et al., 2008; Hashimoto and Panchenko, 2010; Gupta, 2016b; Alnajar et al., 2017).

In addition to the role of surface exposed loops in ligand and protein-protein binding, they can also play an essential role as anchor to help peripheral protein binding or sensing the membrane bilayer surfaces (Kalli and Sansom, 2014; Kalli et al., 2014; Chavent et al., 2016; Xu et al., 2016; Khadka and Gupta, 2017). Recent studies on number of different CSIs in several essential proteins such as DNA Gyrase B, DNA

dependent RNA polymerase Beta Subunit (RpoB), Phosphatidylinositol-4-Phosphate-5-Kinase (PIP5K), Phosphoketolase (PK), Ribonucleotide Reductase (RNR), that are specific for a number of bacterial phyla support the view that these CSIs are involved in conferring ancillary functions on these proteins that are expected to be vital for the organism in which they are found (Griffiths and Gupta, 2004b; Schoeffler et al., 2010; Alnajar et al., 2017; Gupta et al., 2017; Khadka and Gupta, 2017).

Moreover, the structural difference created by CSIs on the surface of protein could act as a unique site that can provide a high-affinity binding site for a specific substrate, ligands, peptides or drug molecules (Cherkasov et al., 2005; Nandan et al., 2007). A preliminary application of CSIs in designing drug compounds has been shown by *Nandan et al.* (2007), who identified a compound that targeted a 12 aa residues deletion in the Elongation factor-1 alpha (EF-1α) of the protozoan parasite *Leishmania donovani*. This deletion provides a unique binding site on the surface of *L. donovani* EF-1α, thereby allowing selective targeting by drug compounds that exhibit greater inhibition of the target *Leishmania* protein than the human homologs (Nandan et al., 2007). Therefore, CSIs could also represent a previously underutilized class of drug targets and should prove useful in the development of novel classes of therapeutics that would selectively target a specific group of pathogenic organisms (e.g. *Mycobacterium tuberculosis*) in which the CSIs are present (Gupta, 2018).

7. Conserved insertions and deletions (Indels) in Protein Structure and Understanding their Function from Bioinformatics and Computational Perspective

An insertion or deletion (indel) in protein sequences represents an important source of genetic variations that shapes an evolution of a protein (Pascarella and Argos, 1992; Benner et al., 1993). However, the mechanisms of indel evolution and their influence on protein structures have been relatively understudied. An indel event in a given protein not only alters the length of its primary sequence but also brings changes within the region of protein domains in which they occur. This, in turn, can influence the structural features and interactions associated with protein region or domain and as a consequence may impair or improve its stability or functions (Pascarella and Argos, 1992; Overington et al., 1992; Benner et al., 1993; Matsuura et al., 1999; Chow et al., 2003; Akiva et al., 2008; Hashimoto and Panchenko, 2010). Indels rarely change the core of a protein structure and are most often found in loops and turns as in those locations they are less likely to disrupt the core or the fold of a protein. The likelihood of occurrence of indels on the surface of the protein is also due to the fact that residues proximity to the core functions are known to impose additional functional constraints on amino acid substitutions or properties (Russell et al., 1997; Chelliah et al., 2004; Jack et al., 2016). The occurrence of indels in the outer surface region in protein structures is considered responsible for the functional divergence of homologous proteins (Reeves et al., 2006; Hashimoto and Panchenko, 2010). Previous studies have shown that vast majorities of protein indels are short typically with a size ranging from 1-6 aa residues in length and mostly occur in surface loops (Pascarella and Argos, 1992; Hsing and

Cherkasov, 2008; Ajawatanawong and Baldauf, 2013). Large indels, however, are most often found to constitute separate secondary structure features forming the regulatory domain (Griffiths and Gupta, 2004b; Chlenov et al., 2005; Hashimoto and Panchenko, 2010; Schoeffler et al., 2010; Gao and Gupta, 2012; Alnajar et al., 2017).

The increasing availability of sequence and structural data have allowed large-scale analyses of the insertions and deletions in proteins using sequence or structural alignment methods and to infer the influence of indels on protein structure or to understand the evolutionary relationships (Hsing and Cherkasov, 2008; Kim and Guo, 2010; Zhang et al., 2011; Zhang et al., 2012). For instance, if the structural information of closely related homologous proteins in alignment is available then by analyzing the alignments or by creating homology models it is possible to infer or "map" the features of indels from the alignment or protein models. Using this approach it is possible to investigate how these genetic changes in the form of indels influence the secondary structure or surface accessibility within the protein structure. It is important to note that, albeit some protein structures, which are useful for structural analyses or homology modelling, may diverge from the physiological reality, most of the available crystal structures offer significant and applicable knowledge about the proteins. Therefore, these kinds of sequence-structure analysis approach has been useful for gaining functional clues for a large number of sequenced proteins that contain indels and for which the structural or experimental information are not yet available (Kristensen et al., 2008; Erdin et al., 2010; Furnham et al., 2012; Furnham et al., 2016).

In the past, some authors have reported databases that provide information about different randomly distributed indels in various proteins (Hsing and Cherkasov, 2008; Zhang et al., 2011; Zhang et al., 2012; Hsing and Cherkasov, 2008). However, unlike the CSIs, most of the indels in those databases are not lineage-specific and are also not present in conserved regions. Thus, these indels are less likely to possess any evolutionary or broad functional significance. In contrast, CSIs are found in conserved regions of the proteins and are unique for the group of organisms in which they are found. Because of their uniqueness and their presence in highly conserved regions of proteins, it is reasonable to infer that most of these CSIs are providing some ancillary function to the proteins or organisms in which they are found. Novel structural alterations in protein structure resulting from the CSIs may contribute to dimer stability, ligand-binding or protein-protein interactions and thus are predicted to confer new characteristics unique to specific group of organisms (Griffiths and Gupta, 2004b; Chlenov et al., 2005; Akiva et al., 2008; Hashimoto and Panchenko, 2010; Schoeffler et al., 2010; Gao and Gupta, 2012; Alnajar et al., 2017; Khadka and Gupta, 2017). Despite their predicted significance, the functional significance of most of the CSIs that modify a part of protein structure remains relatively understudied. An understanding of the peculiarities of these CSIs can, therefore, provide novel insights into the mechanism of protein evolution and as well the ancillary function associated with these CSIs.

During the last two decades, global structural genomics efforts have resulted in the development of powerful new *in vitro* and *in silico* technologies and high-throughput pipelines that are capable of generating hundreds of different protein structures annually

from diverse organism sources (Chance et al., 2004). In addition to the remarkable growth in the sequence and structural data combined with other biological information, the tremendous improvement in computational power and resources have helped build the stronger foundation for the implementation of various computational methods for interpolation between hypothesis and experimental results. As a result, computational approaches utilizing sequence, structural and functional data are now well suited and playing an increasingly important role for studying a wide range of computational experiments, such as mapping the structural features and predicting the functional significance of CSIs in protein structure (Kinoshita and Nakamura, 2003; Watson et al., 2005; Lee et al., 2007; Gherardini and Helmer-Citterich, 2008; Alnajar et al., 2017; Gupta et al., 2017; Khadka and Gupta, 2017). A summary of the major computational tools and techniques that I have utilized throughout my thesis work are briefly described in the following sections. These include BLAST/PSI-BLAST, homology modelling, protein-ligand docking, protein-protein docking, and molecular dynamics simulations.

7.1 BLAST and PSI-BLAST

BLAST (Basic Local Alignment Search Tool) is the most widely used and cited sequence similarity search algorithm; it provides a simple and robust method for searching nucleotide or protein sequence databases to identify significant matches (Altschul et al., 1990). BLAST works by seeking near perfect word-matches "query words" of a given length between the query and database of sequence. If the matches score above a given threshold, the comparison is extended in both directions. Finally, hits in the search results are reported if the extended alignments meet or exceed the user-

specified BLAST cut-offs scores for significant matches. Although BLAST enables searches of large databases for a similar sequence in a relatively short time, it has limited sensitivity in detecting distant homologs.

The PSI-BLAST (Positions-Specific Iterative BLAST), which is an extension of the original BLAST, is more sensitive in searching a database for more evolutionary divergent proteins (Altschul et al., 1997). Iterative search methods such as PSI-BLAST initially generate Position-specific Scoring Matrix (PSSM) constructed from multiple sequences by utilizing the multiple alignment results from normal BLAST. The PSSM stores the conservation pattern for each position in the alignment as a matrix of scores by providing high to low scores relative to their conservation. The new profile is then used to further search the database and this process is then iterated until no additional sequences are found above the predefined threshold. This iterative process makes PSI-BLAST more sensitive in finding remotely related sequences. Due to its effectiveness in finding distantly related sequence, several modifications and improvement have been proposed since its first release in 1997, including employing composition-based statistics (Schaffer et al., 2001), employing pesudocount (Altschul et al., 2009), optimizing cache utilization (Aspnas et al., 2010) and improving sequence weighting method (Oda et al., 2017).

I have utilized both BLAST and PSI-BLAST extensively throughout my thesis work for tasks ranging from identifying homologs/orthologs for species distribution analysis to identifying homologs with known structures for studying the structural features of CSI-containing proteins that lack structural information.

7.2. Homology Modelling for the Prediction of Protein Structures

Over the past few decades, advances in genome sequencing have led to a massive increase in the number of protein sequence data (McPherson, 2014). At the same time, the numbers of protein structures solved by experimental methods lag far behind. As a result, there is a huge and growing gap between the known protein sequences and solved protein structures (Mistry et al., 2013; Rose et al., 2015). In the absence of atomic-resolution experimental structures, exploring the structural features and functional significance of a large number of CSIs in various proteins remains a crucial challenge. Computational approaches such as homology modelling, using an alignment of a novel sequence to that of a sequence with a known protein structure to infer novel structural features, are playing an increasingly important role to bridge this sequence-structure gap (Baker and Sali, 2001; Jaroszewski, 2009; Schwede, 2013).

One of the most commonly used tools for homology modelling is MODELLER (Sali and Blundell, 1993). It is a command line-based tool, which uses python for its control language and requires all its input scripts as python scripts. The process of model generation using MODELLER requires setting up of input files and editing of python scripts for its different steps, which can be both time-consuming and has the potential to introduce errors when working with numerous models. In Chapter 6 of this thesis, I describe the development of a software pipeline to streamline the creation of homology models of proteins using MODELLER and other related programs. I have utilized this pipeline in order to determine and analyze the structural features of a number of CSI-containing proteins without structural information. Structural studies using homology

models have provided a wealth of information to help understand structure-functional relationships of a number of CSIs identified in various proteins (Gao and Gupta, 2012; Gupta et al., 2017; Alnajar et al., 2017; Khadka and Gupta, 2017; Khadka et al., 2017; Hassan and Gupta, 2018; Khadka and Gupta, 2019). The utility of homology modelling to unravel the structural features of a large number of CSIs identified in various proteins will only continue to grow with the increasing experimental structural knowledge from a wide range of protein families. The significance of the integrated pipeline "*GlabModeller*" and its application to study the structural features on several identified CSIs are discussed in detail in Chapter 6 of this thesis.

7.3. Protein-ligand Docking

Protein-ligand docking, first reported in the early 1980s, is a computational approach that aims to predict binding modes of protein-ligand complexes (Kuntz et al., 1982). A docking process comprises two major components, a search algorithm to investigate the possible binding conformations of a ligand in target protein, and an energy scoring functions to evaluate and rank the quality of generated conformations using scoring functions (Hoffmann et al., 1999; Halperin et al., 2002; Meng et al., 2011). The scoring functions utilized by most docking programs can be broadly classified into three categories: (i) Empirical scoring function, approximates protein-ligand binding by adding up the individual weighted terms each representing a key energetic factor in protein-ligand binding (Bohm, 1994; Friesner et al., 2004; Liu and Wang, 2015) (ii) Knowledge-based scoring function, which is developed with training sets of high-resolution structures and searches protein-ligand complexes with an optimal score. It is designed to replicate

experimental structural complexes rather than binding energies (Gohlke et al., 2000; Muegge, 2002) (iii) Force field-based or Physics-based, utilizes the force field parameters to calculate the binding energy of protein-ligand complexes (Huang and Jacobson, 2007).

A great degree of progress has been made over the last few decades in protein-ligand docking programs which exhibit a different level of accuracy and computational efficiencies. Among the several available programs, AutoDock tools (Goodsell and Olson, 1990; Morris et al., 2009) and AutoDock Vina (Trott and Olson, 2010) are arguably the two most popular and highly successful docking tools for protein-ligand docking studies available today. Both these tools are open source and maintained at the Scripps Research Institute (Trott and Olson, 2010). Due to their usefulness in screening libraries of millions of compounds, such tools are becoming an increasing source of lead molecules for drug discovery (McInnes, 2007). Despite the progress, protein-ligand docking still faces many computational challenges, specifically when considering the ligand or receptor flexibility (Sousa et al., 2006). Most docking methods employ rigid docking, where the ligand is treated flexible and receptors are considered rigid, in order to reduce the potential conformational search space to save computational cost and time. Although this allows protein-ligand docking to be carried out at a significantly faster pace, it is considered to be less accurate when compared to the flexible docking. It is therefore important to consider the possible effects of these limitations when employing a protein-ligand docking approach. Nevertheless, due to the significance of these methods to provide an understanding of the key interactions made by ligands or to analyze the potentially novel interaction between small molecules and target site in the protein

structure, they have become extremely desirable tool for studying the possible functional role of CSIs in protein structure.

7.4. Protein-protein Docking

Protein-protein docking is a computational approach that aims to predict the binding conformation of macromolecular complex starting from the individual structure of the component proteins (Levinthal et al., 1975; Wodak and Janin, 1978). Similar to protein-ligand docking, the docking process is composed of two major steps. First, a rigid-body search to sample the possible binding conformation with favorable complementary surfaces and suitable electrostatic and desolvation properties. Second, a refinement step where a suitable scoring function is employed to rank the sampled conformation (Vajda and Camacho, 2004). As with protein-ligand docking, flexibility still constitutes a vital challenge in protein-protein docking due to large computational time that is needed and also due to the lack of sophisticated docking algorithms that allow better treatment of flexibility (Zhang et al., 2016b). Despite these limitations, protein-protein docking methods have improved substantially in recent years and several web-based powerful docking programs are now available to provide an efficient and powerful means for large scale protein-protein docking experiments (Ritchie, 2008; Janin, 2010; Moreira et al., 2010; Lensink et al., 2017). Among these, the tools that are widely used for computational protein-protein studies and that I have utilized during my thesis work includes, PatchDock (Schneidman-Duhovny et al., 2005), ZDOCK (Pierce et al., 2011), ClusPro (Comeau et al., 2004) and RosettaDock (ROSIE) (Lyskov et al., 2013).

PatchDock is an efficient molecular docking algorithm that employs a geometry-based shape complementarity approach which aims to yield refined atomic contacts of protein-protein complexes. It's scoring function takes into consideration both geometric fit and atomic desolvation energy (Schneidman-Duhovny et al., 2005). ZDOCK utilizes grid-based fast Fourier transform (FTT) for efficient global search of docking orientation between two proteins (Pierce et al., 2011). Its scoring function is based on pairwise shape complementarity, electrostatics, and a pairwise atomic statistical potential developed using contact propensities of transient protein complexes (Chen and Weng, 2003; Pierce et al., 2011). ClusPro utilizes PIPER, a rigid body docking program, which is based on a novel Fast-Fourier Transform (FFT) docking approach with pairwise potential. Its scoring function is thus based on pairwise interaction potentials (Comeau et al., 2004; Kozakov et al., 2006). RosettaDock (ROSIE), utilizes a Monte Carlo-based algorithm, to search for the rigid-body and side-chain conformational space of two interacting macromolecules and finds their structure complex with minimum free-energy (Gray et al., 2003; Lyskov and Gray, 2008; Lyskov et al., 2013). These protein-protein docking servers provide a platform to carry out extremely fast docking experiments without requiring significant computational time and resources. Further, the applicability and utility of protein-protein docking will continue to grow as the structural genomics initiatives continue to populate the space of 3D structures knowledge of many cellular complexes. In most cases, it is likely that many such structures will contain in their structure these unique molecular markers (CSIs), and an efficacious protein-protein docking approach in combination with other computational approaches such as homology modelling, would provide a path

toward understanding the functional significance of large number of identified CSIs in protein structures (Alnajar et al., 2017; Gupta et al., 2017). As an example, in Chapter 5 of this thesis, I describe the use of efficacious protein-protein docking approaches in predicting the functional role of CSIs in the Phosphoketolase proteins.

7.5. Molecular Dynamics Simulations

In addition to the knowledge of 3D structure of a protein, a fundamental appreciation for insights into the protein function requires understanding the relationship between 3D structure of a protein and its dynamics (Levitt and Warshel, 1975; McCammon et al., 1977; Karplus and Kuriyan, 2005). Computational methods such as molecular dynamics (MD) simulations play an increasingly important role by providing a link between protein structure and dynamics (Karplus, 2002; Karplus and Kuriyan, 2005; Freddolino and Schulten, 2009; Klepeis et al., 2009; Bermudez et al., 2016; Hollingsworth and Dror, 2018). In my thesis work I have utilized, in particular, a classical molecular dynamics simulation known as a "*computational microscope*" (Lee et al., 2009) to investigate the molecular interaction of biomolecular systems with an aim to shed light into the structural and functional significance of CSIs.

The basic principle behind classical MD simulations is to predict the behavior of an individual atom in a protein or other molecular systems using the physics governing interatomic interactions over the function of time (Karplus, 2002). The concept is that, given the position of all the atoms in a given biomolecular system, which may involve protein solvated by waters molecules or embedded in a lipid bilayer membrane model, it is possible to estimate the force excreted on each atom by all other atoms in the system

using Newton's second law or equation of motion. Based on the knowledge of those forces, spatial position or acceleration and velocity of each atom in a system as a function of time can then be calculated. The trajectories thus obtained can be used to generate astonishingly detailed 3D representations of how a biomolecule behaves over time under a variety of tunable conditions. Recent extensive comparison of a variety of experimental data with simulations suggests that force fields have improved significantly over time but certain deficiencies still persist requiring future improvement (Lindorff-Larsen et al., 2012). Therefore when analyzing the results from the MD approach, it is important to consider the possible effects of these limitations (Lindorff-Larsen et al., 2012; Poger et al., 2016; Nerenberg and Head-Gordon, 2018).

Cognizant of these limitations, MD simulations have greatly expanded the scope of biomolecular science, drug discovery and several other fields of science. It has been used successfully to study the conformational changes of a number of different proteins, which otherwise would have been difficult or impossible to determine by experimental techniques (Ma et al., 2000; Holyoake and Sansom, 2007; Freddolino and Schulten, 2009; Klepeis et al., 2009; Bermudez et al., 2016; Hall et al., 2019). Specifically, membrane-associated proteins (i.e. both integral and peripheral), for which the structural information is very scarce due to barriers associated with their expression and crystallization (Vattulainen and Rog, 2011; Biggin and Bond, 2015) have greatly benefited from the MD approach. This growing application potential have led to the implementation of MD simulation in several software packages which include AMBER (Case et al., 2005), CHARMM (Bernard et al., 1983), DESMOND (Shivakumar et al., 2010), GROMACS

(Van Der et al., 2005), LAMPS (Plimpton, 1995) and NAMD (Phillips et al., 2005). Among these programs, GROMACS (GROningen MAchine for Chemical Simulations) stands out from others due to its efficiency, using an ensemble method to make the best possible usage of scarce computational resources (Abrahama et al., 2016).

Over the last few years, substantial progress has been made in computer hardware, specifically the graphics processing units (GPUs), which have fueled the remarkable improvement in speed and accuracy of computing software and MD algorithms. These improvements now allows powerful simulations to be run locally for longer time at a very affordable cost (Klepeis et al., 2009; Friedrichs et al., 2009; Luttmann et al., 2009; Mashimo et al., 2013; Kutzner et al., 2015; Hollingsworth and Dror, 2018; Larsson et al., 2019). It is important to note that, although a longer time simulation run provides high-resolution atomic details of macromolecular behaviour, however, they are computationally too intensive. As a result, for the MD simulation studies that I have described in Chapter 3 of this thesis, I have utilized a very reasonable length of simulation which makes comprehensive simulation studies very feasible. Nonetheless, the cumulative efficiency and availability make MD simulation particularly useful for analyzing the large number of CSIs and to generate and test a mechanistic hypothesis of how these CSIs function in different proteins. As most of the CSIs are found to be located on the surface loop of the proteins, the analysis of the loops in terms of their conformational changes during MD simulation is vital to unravel their potential function. For instance, analyses of an array of binding events during protein-ligand and/or protein-protein or protein-membrane interaction processes, the role of water molecules. and also

the effects of mutations and the modification of protein ensembles and their functions (Gu et al., 2009; Dror et al., 2011; Buch et al., 2011; Gu et al., 2015; Bermudez et al., 2016; Khadka and Gupta, 2017; Wacker et al., 2017; Rudling et al., 2018).

8. Research Objectives

The major focus of my graduate research work has been two-fold: 1) identification and analysis of CSIs which are distinctive characteristics of fungi as well as of the multiple important groups of microorganisms, and utilization of phylogenomic and comparative genomic approaches to elucidate their evolutionary history and relationships. 2) utilization of various computational and bioinformatic approaches to investigate the identified CSIs present within various essential proteins in order to elucidate their structural and functional significance.

With the aim of understanding the evolutionary significance, I have utilized a combination of phylogenetic analysis and CSI identification to provide important insight into the evolutionary history of the PIP4K/PIP5K family of proteins. The members of PIP4K/PIP5K family are key players in the regulation of the metabolism of phosphatidylinositides (PI), which act as a secondary messenger for controlling diverse cellular processes in eukaryotes (Majerus, 1992; Martin, 1998; Di Paolo and De Camilli, 2006; van den Bout and Divecha, 2009; Kutateladze, 2010; Epand, 2017). This family of proteins, that shares sequence identity within the kinase domain, are classified into three distantly related groups of proteins viz. Phosphatidylinositol-4-phosphate 5-kinase Type I (PIP5K), phosphatidylinositol-5-phosphate-4-kinase Type II (PIP4K), and

phosphatidylinositol-3-phosphate 5-kinase Type III (or PIKFYVE) (Loijens et al., 1996; Heck et al., 2007; Brown and Auger, 2011). There are three different isoforms identified for both PIP5K (PIP5Kα, PIP4Kβ, and PIP4Kγ), and PIP4K (PIP4Kα, PIP4Kβ, and PIP4Kγ) in vertebrates (Ishihara et al., 1996; Ishihara et al., 1998). However, invertebrates have been reported to contain only a single homolog of both these proteins (Brown and Auger, 2011). Similarly, in fungi, a single copy homologs of PIP5Ks are found that shows similarity to both PIP4K and PIP5K (Desrivieres et al., 1998), whereas in plants multiple homologs that show similarity to both PIP5K and PIP4K are present (Okazaki et al., 2015; Heilmann, 2016). Due to the important roles played by the PIP4K/PIP5K family of proteins in many critical processes involved in pathological conditions these proteins are becoming an increasingly interesting class of molecular targets for cancer (Emerling et al., 2013; Semenas et al., 2014), chronic pain (Wright et al., 2015), diabetes (Voss et al., 2014), and autoimmune diseases (Hayakawa et al., 2014). Despite their important roles in the regulation of many cellular processes, our understanding of the overall evolutionary relationships between and among different members of the PIP4K/PIP5K families and subfamilies of proteins and the presence of any unique genetic/biochemical characteristics that can serve to distinguish different members of this protein family remains largely enigmatic and unexplored. In my subsequent work to investigate the functional differentiation of the different members of the PIP4K/PIP5K family, I have utilized one of the CSIs identified in the PIP4K/PIP5K family of proteins that is specific for the *Saccharomycetaceae* family of fungi.

To further understand the functional significance of CSIs, I have also studied the Phosphoketolase (PKs) enzyme from Bifidobacteria, which constitute an important group of commensal bacteria that inhabit the gastrointestinal tracts of humans, other mammals, as well as insects (Biavati et al., 2000; Turroni et al., 2011; Ventura et al., 2014). Bifidobacteria are known to exert several health-promoting benefits on their host (Pokusaeva et al., 2011; Sanchez et al., 2017). One important characteristic of bifidobacteria is the presence of a unique fermentation pathway known as the "bifid shunt" for the metabolism of different carbohydrates (Meile et al., 2001; Takahashi et al., 2010). The key enzyme involved in this pathway is phosphoketolase (PKs) and unlike phosphoketolase (XPKs) from other bacteria, which shows specificity only for only Xylulose-5-phosphate (X5P), the bifidobacteria phosphoketolase (XFPK) possess an unique ability to metabolize both X5P and Fructose 6-phosphate (F6P) (Meile et al., 2001; Yin et al., 2005; Takahashi et al., 2010; Henard et al., 2015). Despite the well-known differences in the biological activities of PKs between bifidobacteria and other bacteria, very little is known about the molecular or biochemical and/or structural characteristics accounting for the important differences in the two forms of PKs.

I have also extended my analysis of the structural and functional significance of CSIs to the CSIs identified in the SecA proteins that are unique to some thermophilic and hyperthermophilic group of bacteria. SecA, a conserved ATPase, is a multifunctional dynamic protein that forms a key component of bacterial Sec-translocation system (Mori and Ito, 2001; Vrontou and Economou, 2004). SecA is essential for the survival of broad-spectrum bacteria, as well as archaea, and plays an indispensable role in the secretion of a

wide variety of bacterial proteins (Schmidt and Kiser, 1999; Gil et al., 2004). Earlier studies using comparative genomic analysis had led to the identification of a large number of CSIs that served as a unique molecular characteristics for the orders *Thermotogales*, *Acquificales* and *Thermales* (Griffiths and Gupta, 2006; Gupta and Bhandari, 2011) which contains some of the most hyperthermophilic species of bacteria known to date, with an upper temperature limit of growth up to 95°C (Vieille and Zeikus, 2001). Although previous biochemical and structural studies have contributed significantly towards understanding the overall architecture and function of the SecA protein (Zimmer and Rapoport, 2009; Chen et al., 2015; Milenkovic and Bondar, 2016). It remains yet unclear whether the unique presence of these large CSIs contributes to the stability of SecA or it may be important for the SecA to function at high temperatures in thermophilic bacteria.

9. Outline of this Thesis

The analysis completed in my thesis work has provided novel insights into the understanding of evolutionary significance as well as the unique structural and functional aspect of several CSIs in key proteins involved in essential pathways in different organisms (Gupta et al., 2017; Khadka and Gupta, 2017; Khadka and Gupta, 2019). In Chapter 2 of this thesis, I provide novel insights into the origin, evolutionary relationship, and diversification of PIP4K/PIP5K protein family. In this chapter, I described in-depth analyses of species distribution and have carried out detailed phylogenetic studies and comparative analyses of protein sequences to identify many molecular markers in the

form of CSIs that are specific for the different members of PIP4K/PIP5K family of isoenzymes. Our CSI-based approach in conjunction with the results obtained from the BLASTp searches for the distribution of these isoenzymes provides novel insights into the evolutionary history of PIP4K/PIP5K family of protein. In Chapter 3 of this thesis, I describe our subsequent work on the identification and analysis of an 8 aa CSI, specific for the *Saccharomycetaceae* family of fungi. Here, I describe the analysis of this CSI present in a core conserved region of PIP5K, a key enzyme in the phosphatidylinositol signaling pathway essential for multiple cellular processes (Di Paolo and De Camilli, 2006; Balla et al., 2009; Balla, 2013; Epand, 2017). Based on our results from structural analysis and molecular dynamics (MD) simulation studies, we provided useful insights concerning the mechanism of the interaction of PIP5K with lipid bilayer and support the idea that the 8 aa CSI in *S. cerevisiae* plays an important role in facilitating the binding of PIP5K with a membrane surface.

In Chapter 4, I describe the identification of multiple highly specific molecular differences in the form of CSIs that clearly distinguish the phosphoketolase of bifidobacteria from the phosphoketolase homologs found in most other bacteria. We also provide evidence, based on the analyses of the branching pattern from phylogenetic tree, that the PKs in bifidobacteria (XFPK) are specifically related to those found in the *Coriobacteriales*, indicating that the gene for this protein was horizontally transferred between these two groups. Additionally, we also describe in this chapter the utilization of molecular modelling, structural analyses and protein-protein docking studies to unravel that the *Bifidobacteriales*/*Coriobacteriales* specific CSIs are located on the surface

exposed loop region at the subunit interface in the XFPK structure and that they are involved in the formation/stabilization of XFPK dimer.

Chapter 5 of this thesis describes the identification and analysis of several large CSIs in SecA proteins that are uniquely shared by the members of the order *Thermotogales*, *Aquificales,* and *Thermales*, which represent the major thermophilic phyla. In this chapter, I describe the sequences and phylogenetic analyses of these proteins which provide suggestive evidence for convergent evolution resulting in the origin of the insertions in these distantly related groups and were likely retain due to their selective advantageous functional roles. To further unravel the functional significance of thermophilic and hyperthermophilic specific CSIs, I explore the molecular dynamics (MD) simulations using the *Thermotoga maritima* SecA structure with and without CSI at various temperature setting. The results from MD studies identified a conserved network of water molecules and conserved residues within the CSI that make key contributions toward these interactions. Chapter 6 in this thesis describes an interactive graphical pipeline program called "*GlabModeller*" which provides an easy-to-use graphical user interface (GUI) for Modeller, a homology modelling program, and a number of subsequent steps involved in the model refinement and validation process. The utility of this tool to study the mapping of the structural location and structural features of various CSIs is evident in Chapter 2, Chapter 3, Chapter 4 and Chapter 5 of this thesis. Finally, Chapter 7 describes the overall significance of the evolutionary, structural and functional studies on CSIs described in this thesis, usefulness of the integrated pipeline program, and the potential future direction for this work.

CHAPTER 2

Novel Molecular Signatures in the PIP4K/PIP5K Family of Proteins Specific for Different Isozymes and Subfamilies Provide Important Insights into the Evolutionary Divergence of this Protein Family

This chapter describes the applications of the comparative genomics approach for the identification of CSIs unique to the members of the PIP4K/PIP5K family of proteins. The identified CSIs in conjunction with species distributions and phylogenetic analyses based on their protein sequences shed light on the origin and evolutionary history of this protein family. This chapter also describes the mapped locations and structural features of all identified CSIs onto the structure of PIP4K and PIP5K proteins. My contribution toward the completion of this chapter includes identification of all the CSIs shown and confirmation of their species specificities, analyses of the species distribution of PIP4K/PIP5K homologs, construction of the phylogenetic tree, generation of the homology models and the analyses of the structural features and localization of the identified CSIs in the protein structures, the writing of drafts and revision of the manuscript, and production of all main and supplemental figures and tables in the manuscript.

Due to limited space, supplementary materials (figures and tables) are not included in the chapter but can be accessed along with the rest of the manuscript at:

Khadka, B., & Gupta, R. S. (2019). *Genes*, 10(4), 312.

genes

MDPI

Article

Novel Molecular Signatures in the PIP4K/PIP5K Family of Proteins Specific for Different Isozymes and Subfamilies Provide Important Insights into the Evolutionary Divergence of this Protein Family

Bijendra Khadka and Radhey S. Gupta *

Department of Biochemistry and Biomedical Sciences McMaster University, Hamilton, ON L8N 3Z5, Canada; khadkab@mcmaster.ca

* Correspondence: gupta@mcmaster.ca; Tel.: +1-905-525-9140

Received: 22 February 2019; Accepted: 15 April 2019; Published: 21 April 2019

Abstract: Members of the PIP4K/PIP5K family of proteins, which generate the highly important secondary messenger phosphatidylinositol-4,5-bisphosphate, play central roles in regulating diverse signaling pathways. In eukaryotic organisms, multiple isozymes and subfamilies of PIP4K/PIP5K proteins are found and it is of much interest to understand their evolution and species distribution and what unique molecular and biochemical characteristics distinguish specific isozymes and subfamilies of proteins. We report here the species distribution of different PIP4K/PIP5K family of proteins in eukaryotic organisms and phylogenetic analysis based on their protein sequences. Our results indicate that the distinct homologs of both PIP4K and PIP5K are found in different organisms belonging to the Holozoa clade of eukaryotes, which comprises of various metazoan phyla as well as their close unicellular relatives Choanoflagellates and Filasterea. In contrast, the deeper-branching eukaryotic lineages, as well as plants and fungi, contain only a single homolog of the PIP4K/PIP5K proteins. In parallel, our comparative analyses of PIP4K/PIP5K protein sequences have identified six highly-specific molecular markers consisting of conserved signature indels (CSIs) that are uniquely shared by either the PIP4K or PIP5K proteins, or both, or specific subfamilies of these proteins. Of these molecular markers, 2 CSIs are distinctive characteristics of all PIP4K homologs, 1 CSI distinguishes the PIP4K and PIP5K homologs from the Holozoa clade of species from the ancestral form of PIP4K/PIP5K found in deeper-branching eukaryotic lineages. The remaining three CSIs are specific for the PIP5Kα, PIP5Kβ, and PIP4Kγ subfamilies of proteins from vertebrate species. These molecular markers provide important means for distinguishing different PIP4K/PIP5K isozymes as well as some of their subfamilies. In addition, the distribution patterns of these markers in different isozymes provide important insights into the evolutionary divergence of PIP4K/PIP5K proteins. Our results support the view that the Holozoa clade of eukaryotic organisms shared a common ancestor exclusive of the other eukaryotic lineages and that the initial gene duplication event leading to the divergence of distinct types of PIP4K and PIP5K homologs occurred in a common ancestor of this clade. Based on the results gleaned from different studies presented here, a model for the evolutionary divergence of the PIP4K/PIP5K family of proteins is presented.

Keywords: phosphatidylinositol phosphate kinases; conserved signature indels; molecular signatures for the pip4k/pip5k isozymes and isoforms; phylogenetic analysis; species distribution of pip4k/pip5k proteins; holozoa clade of eukaryotic organisms; evolution of the PIP4K/PIP5K family of proteins; protein evolution

1. Introduction

The members of the PIP4K/PIP5K family of proteins play key roles in the synthesis and regulation of various phosphoinositides (PIs), which act as a secondary messenger for controlling diverse cellular processes in eukaryotes [1–3]. These proteins (or isozymes) are classified into three distantly related groups of proteins viz. (i) Phosphatidylinositol-4-phosphate 5-kinase Type I (PIP5K), (ii) phosphatidylinositol-5-phosphate 4-kinase Type II (PIP4K), and (iii) phosphatidylinositol-3-phosphate 5-kinase Type III (or PIKFYVE) [4,5]. The PIP5K family of isozymes, which catalyze the conversion of phosphatidylinositol-4-phosphate (PI4P) to phosphatidylinositol-4,5-bisphosphate ($PI(4,5)P_2$), are mainly localized to the plasma membrane, Golgi complex, and nucleus in mammalian cells [6,7]. In vertebrates, three subfamilies or isoforms (encoded by separate genes) of PIP5K (viz. PIP5Kα, PIP5Kβ, and PIP5Kγ) are present along with multiple splice variants of these isoforms generated as a result of alternative splicing [8]. For instance, in humans, three splice variants of PIP5Kα, four splice variants of PIP5Kβ, and three splice variants of PIP5Kγ have been reported [5,8,9]. The PIP4K on the other hand, which catalyzes the phosphorylation of the 4-hydroxyl group of phosphatidylinositol-5-phosphate (PI5P), are diffusively localized in the cytoplasm, endoplasmic reticulum, actin cytoskeleton, and nucleus [10]. As with the PIP5K, three subfamilies of PIP4K (viz. PIP4Kα, PIP4Kβ, and PIP4Kγ encoded by separate genes) are also found in vertebrates [4,11]. The presence of multiple distinct isoforms of PIP4K/PIP5K in vertebrate species is predicted to enable these organisms to coordinate the regulation of $PI(4,5)P_2$ production for specific processes, either by differential regulation or selective subcellular localization of the individual copies of these enzymes and their isoforms [3,5,8,10,12]. In both PIP5K and PIP4K families, the core kinase domain is highly conserved whereas the regions outside this domain show limited sequence similarity [4,13].

Unlike vertebrates which contain multiples subfamilies of both PIP4K and PIP5K, all invertebrates including Coelomates (Deuterostomes, Protostomata), and Pseudocoelomate have been reported to contain only a single homolog of both these proteins [4]. In fungi, a single homolog of PIP4K/PIP5K showing similarity to both PIP4K and PIP5K is found (referred to as a multiple-copy suppressor of stt4 mutation, MSS4 protein) [14,15]. Plants also contain multiple copies of a protein showing similarity to both PIP4K/PIP5K [16–18]. However, the plants and fungi proteins contain distinct structural features/domains, which are not found in the PIP4K/PIP5K homologs from vertebrates and invertebrates species, suggesting that they perform additional novel functions which are specific for these organisms [17–19]. The PYKFYVE, which catalyzes the phosphorylation of PI3P to $PI(3,5)P_2$ [20], is considerably larger than the PIP4K and PIP5K homologs and it shows very limited similarity to the PIP4K/PIP5K proteins, restricted to the kinase domain.

Due to their important cellular roles, the PIP4K/PIP5K family of isozymes have been studied extensively [21–23]. However, a detailed understanding of the species distribution of these proteins, how the different isozymes and subfamilies of these proteins have evolved and diversified, and novel sequence features that are specific for these two isozymes or their different forms which are present in vertebrate species, remain largely enigmatic and unexplored. In recent years, genome sequences have become available from diverse eukaryotic organisms providing a valuable resource for identifying novel molecular markers/characteristics that are uniquely found in specific proteins from an evolutionarily related group of organisms, or those that are commonly shared by members from a related group/family of proteins. One important class of molecular markers discovered by genome sequence analysis, which has proven very useful for evolutionary, genetic and biochemical studies, is comprised of conserved signature indels (insertions/deletions) (CSIs) in gene/protein sequences [24–29]. The CSIs in gene/protein sequences generally result from rare genetic changes and based upon their presence or absence in different species or homologs, important inferences regarding evolutionary relationships can be derived. In our recent work, multiple CSIs were identified within the catalytic domain of the diacylglycerol kinase (DGK) family of isozymes which were either specific for a particular

class of isozyme or commonly shared by two or more classes of DGK isozymes, thereby providing important insights into the evolutionary history of this protein family [30].

In the present study, we have used a combination of phylogenetic approaches and the CSI identification strategy to understand the evolutionary relationships and the origin/distribution of the PIP4K/PIP5K family of proteins. Our analyses of this protein family have identified six CSIs that are specific characteristics of either all or particular types and subfamilies of the PIP4K/PIP5K proteins revealing their novel sequence features and providing important insights into their evolutionary history. The identified CSIs provide novel tools for functional studies on these proteins and we discuss here the implications of the results presented here for the origin and evolutionary diversification of the PIP4K/PIP5K family of proteins.

2. Materials and Methods

2.1. Identification of Conserved Signature Indels and Phylogenetic Analysis

Protein sequences for the PIP4K/PIP5K family were obtained from the NCBI database [31]. Identification of CSIs (insertions/deletions) in the sequence alignments of these proteins was carried out as described in our earlier work [30,32,33]. In brief, homologs of the PIP4K/PIP5K family proteins from [4,9,13] representative species from major eukaryotic groups were retrieved from the NCBI database and multiple sequence alignments of these proteins were created separately and in combination using the Clustal X program [34]. The α, β, and γ isoforms or subfamilies of the PIP4K and PIP5K proteins present in vertebrates that were analyzed in the present work are encoded for by distinct genes. Although the term isoform is commonly used to refer to these proteins, these proteins are products of separate genes and not derived using alternative splicing. Sequence alignments of the proteins were visually inspected for the presence of CSIs that were flanked on both sides by at least 3–4 conserved amino acids in the neighboring 40–50 amino acids. Indels which were not flanked by conserved regions were not further studied. For all conserved indels thus identified, detailed BLASTp searches were carried out on short sequence segments containing the indel and its flanking conserved regions (60–100 amino acids long) to determine the specificity and species distribution of the indels in PIP4K/PIP5K homologs from different organisms. The CSIs figures shown here were generated using SIG_CREATE and SIG_STYLE programs (from www.GLEANS.net) as described in our earlier work [32,35]. Due to space constraints, sequence information for the PIP4K/PIP5K homologs is shown in the presented figures for only representative species from different groups of organisms. However, unless otherwise indicated, the identified CSIs are specific for the indicated PIP4K/PIP5K isozymes/isoforms for the indicated groups of eukaryotic organisms and detailed information regarding the species distribution of different described CSIs is provided in Figures S2–S7.

For the construction of the phylogenetic tree, sequences for the PIP4K, PIP5K, and PIP4K/PIP5K proteins from different organisms were trimmed to correspond to the core catalytic kinase domain, which is conserved among these homologs. Multiple sequence alignment of the core catalytic kinase domain which consisted of 217 aligned amino acid positions was used for phylogenetic analysis. A maximum-likelihood phylogenetic tree based on 100 bootstrap replicates of this sequence alignment was constructed using MEGA 6 [36] based on the Jones–Taylor–Thornton (JTT) model [37], as detailed in our earlier studies [30,33,38].

2.2. Homology Modelling and Structural Analyses

Homology modeling of the CSI-containing and CSI-lacking PIP5K family of proteins was carried out as described in earlier work using an in-house pipeline, "GlabModeller," for comparative protein structure modeling [33,38,39]. Initially, to identify appropriate templates for homology modeling, PSI-BLAST [40] searches were carried out against the Protein Data Bank (PDB) [41] using protein sequences from the PIP4K/PIP5K homologs. The suitable templates identified by blast searches include, PIP5K-α (*Danio rerio*) (PDB ID: 4TZ7, chain "A"), PIP4K-β (*Homo sapiens*) (PDB ID: 1BO1,

chain "B"), PIP4K-γ (*H. sapiens*) (PDB ID: 2GK9, chain "A"), PIP4K-α (*H. sapiens*) (PDB ID: 2YBX, chain "A") [42,43] and exhibits sequence identities of 71%, 33%, 32%, 30% with the whole sequence of PIP5Kβ (*H. sapiens*) (Accession number: AAH30587.1). For homology modeling, the sequence alignments between target and template proteins were carried out using the align 2D module from the Modeller, which is integrated and streamlined in GlabModeller tool. The resulting alignments obtained were then carefully analyzed and modified manually to ensure the reliability of the location of insertion and deletions. For each target protein, 500 models were generated initially and ranked and selected using the discrete optimized potential energy (DOPE) score [44] as implemented in Modeller v9.15 [45]. The models with high DOPE scores are then submitted to the ModRefiner program to obtain atomic-level energy minimization and to obtain a model with reliable stereochemistry quality [46]. The illustrative approach was utilized to improve the overall quality of the final model. Loop regions are refined using the ModLoop server [47]. The qualities of the final models were analyzed using a number of different independent protein model validation servers/tools which are integrated into the GlabModeller. The servers/tool for validation utilized include RAMPAGE [48], ProSA [49,50], and QMEAN [51]. Visualization and structural analysis of structural models of PIP4K/PIP5K proteins were carried out using the molecular visualization program PyMOL (www.pymol.org).

3. Results

3.1. Species Distribution and Phylogenetic Analysis of PIP4K/PIP5K Protein Family

As noted in the introduction, in contrast to all vertebrates and some invertebrate species (viz. belonging to the superphyla Deutrostomia, Protostomia and Pseudocoelomates), which contain distinct homologs of both PIP4K and PIP5K proteins, plants and fungi contain only a single homolog of these proteins showing similarity to both PIP4K and PIP5K homologs [4]. However, the distribution of these two proteins in other early branching metazoan lineages such as Placozoa, Porifera, Cnidaria, and Ctenophora, or their known sister groups which includes Choanoflagellates and Filasterea, remains undetermined [52–55]. Hence, the distribution of PIP4K and PIP5K isozymes in these early branching metazoan/eukaryotic lineages is of much importance for understanding the evolutionary diversification of PIP4K and PIP5K isozymes within the eukaryotes.

BLASTp searches were carried out with the sequences for PIP4K and PIP5K proteins against the NCBI nr database as well as genome sequences from different eukaryotic organisms. The results from these studies, which are summarized in Table 1, show that distinct homologs of both PIP4K and PIP5K are found in all major metazoan groups (i.e., Bilateria, Cnidaria, Placozoa, and Porifera) as well as in their closest-known unicellular ancestor, Choanoflagellates. Homologs for PIP4K and PIP5K were also found in a Filasterea species (*Capsaspora owczarzaki*), however, no sequence sharing similarity to PIP4K or PIP5K was detected in Ichthyosporea. Outside of the metazoans, Choanoflagellates and Filasterea, only a single homolog of PIP4K/PIP5K was detected in other available genomes from eukaryotic organisms including those from the phyla Apicompelxa, Amebozoa, Percolzoa, and Apusozoa. Further, as known from earlier work [4,11,33], single orthologs exhibiting similarity to both PIP4K and PIP5K were detected in different plants and fungi.

The evolutionary relationship among the PIP4K/PIP5K homologs from different eukaryotic organisms was further investigated by constructing a maximum-likelihood phylogenetic tree based on sequence alignments for the core catalytic domain which is conserved in all members of this protein family. This tree encompasses 96 protein sequences and it includes different PIP4K/PIP5K homologs from representative species of all major taxonomic groups within eukaryotes (Figure 1).

As seen from Figure 1, the PIP4K and PIP5K homologs from different metazoan species form two strongly supported clusters in the tree and they are separated by a long branch from a cluster comprising of the single homologs of PIP4K/PIP5K found in the deeper branching eukaryotic lineages as well as in plants and fungi. However, the overall support for this latter cluster, as well as the interrelationships of different species within it, is generally weak and not reliably resolved. Nonetheless,

based on this phylogenetic tree (Figure 1), several inferences can be drawn: (i) Within the clusters corresponding to PIP4K and PIP5K homologs, although the interrelationships of different metazoans species are not resolved, in both cases the Choanoflagellates and Filasterea species, which are unicellular organisms most closely related to the multicellular metazoans, form the deepest branching lineages in the clusters. (ii) The vertebrate species form strongly supported clades in both PIP4K and PIP5K clusters. The three different families (or types) of PIP4K and PIP5K proteins (viz. α, β and γ) which are found in vertebrate species also formed distinct clades, consistent with earlier studies [4,11]. (iii) Based on their branching in the phylogenetic tree, for PIP4K homologs, the γ-subfamily of the protein exhibited the deepest branching and it formed a sister group of a clade consisting of the PIP4Kα and PIP4Kβ proteins. On the other hand, for the PIP5K protein, the β subfamily showed the earliest divergence followed by the emergence of α and γ paralogs of the proteins. (iv) In the phylogenetic tree, the cluster consisting of the PIP5K homologs exhibited a closer relationship to the cluster comprising of the single copy homologs of PIP4K/PIP5K found in the deeper branching eukaryotic lineages.

Table 1. Distribution of PIP4K/PIP5K family of proteins in the major groups of eukaryotes.

		Taxa/Phylum	PIP4K	PIP5K	PIP4K/PIP5K
DEUTEROSTOMIA	Vertebrates	Mammals	>100	>100	-
		Birds	>100	>100	-
		Amphibians	4	4	-
		Reptiles	10	10	-
		Fishes	>50	>50	-
		Tunicata	2	2	-
		Cephalochordata	2	2	-
		Hemichordata	1	1	-
		Echinodermata	1	1	-
PROTOSTOMIA		Arthropoda	>100	>100	-
		Nematoda	>50	>50	-
		Mollusca	4	4	-
		Annelida	2	2	-
		Platyhelminthes	2	2	-
		Tardigrada	2	2	-
Early Metazoans		PLACOZOA	1	1	-
		PORIFERA	1	1	-
		CNIDARIA	3	3	-
UMCA		CHOANOFLAGELLATEA	2	2	-
		FILASTEREA	1	1	-
		ICHTHYOSPOREA	1 *	1 *	-
		FUNGI	-	-	>100
		PLANTS	-	-	>100
OTHERS EUKARYOTES		APICOMPLEXA	-	-	>10
		AMOEBOZOA	-	-	
		PERCOLOZOA (*Naegleria gruberi*)	-	-	
		APUSOZOA (*Thecamonas trahens*)	-	-	

* 1 hit showing low sequence conservation was observed. The numbers in different columns indicate the number of species from the indicated groups for which sequence information was available.

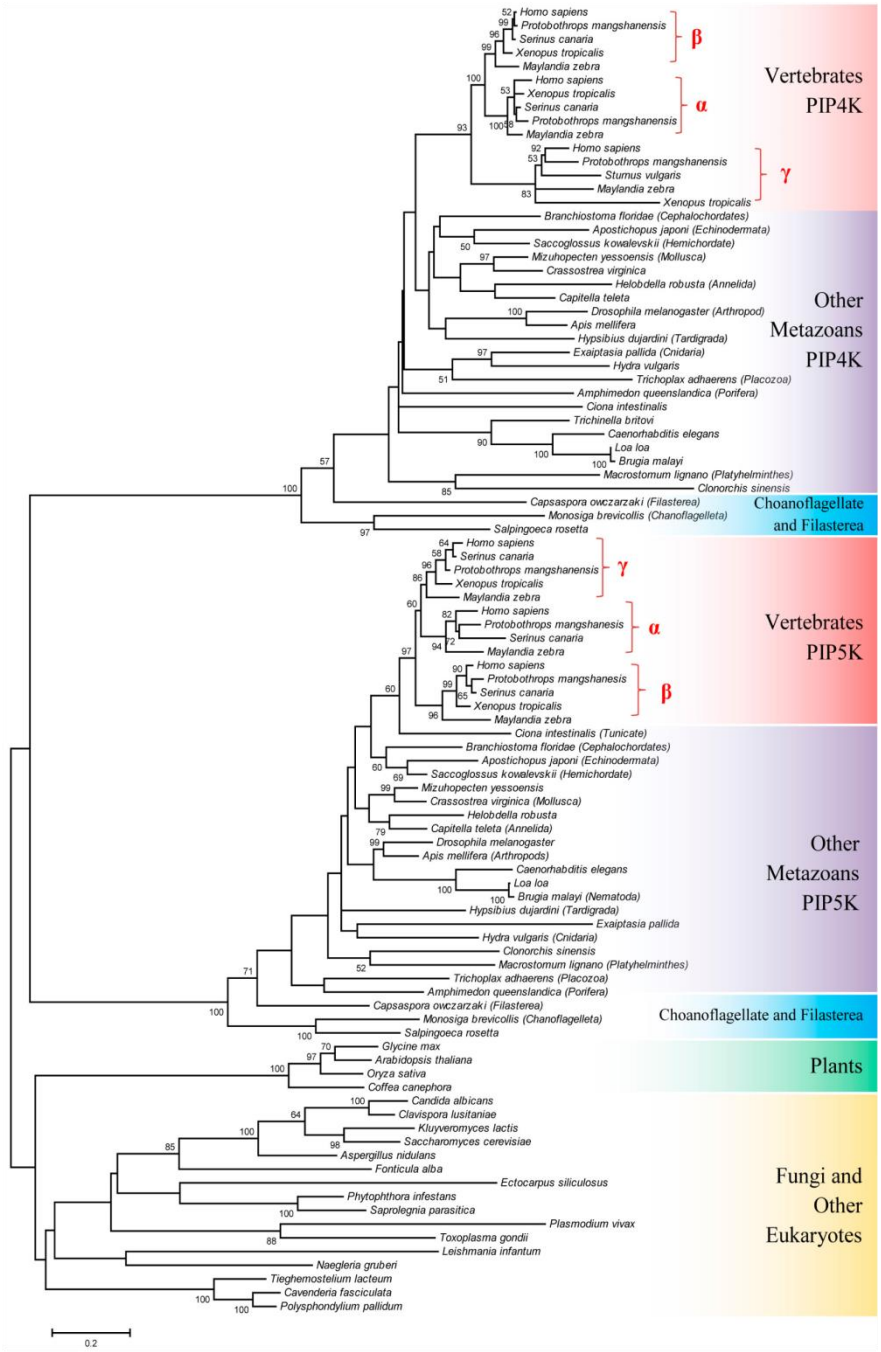


Figure 1. A maximum-likelihood phylogenetic tree of the PIP4K/PIP5K family of proteins based on the core conserved kinase domain region of the protein sequences from representative species. The accession numbers of the protein sequences that were utilized are provided in Table S1. Bootstrap support values > 50% are shown at the nodes.

3.2. Conserved Signature Indels that are Distinctive Features of the PIP4K and PIP5K Family of Proteins and the Insights Provided by Them into the Evolutionary Relationships

Based on the phylogenetic tree in Figure 1 although some inferences regarding evolutionary relationships amongst PIP4K and PIP5K can be drawn, due to poor statistical support and separation by long branches of many significant nodes, it is important to confirm these inferences by other independent approaches. CSIs represent an important class of molecular markers that have been used in the past to resolve a number of important evolutionary questions [24,26,27,29,56]. The CSIs in gene/protein sequences generally result from rare genetic changes. Due to the discrete nature of genetic changes represented by CSIs and their presence in conserved regions, the presence or absence of CSIs in different lineages (or proteins) is generally not affected by factors that can confound branching in phylogenetic trees [26,27,29,56]. In view of the usefulness of CSIs for evolutionary studies, sequence alignments of the PIP4K/PIP5K proteins were examined for the presence of any useful/informative CSIs.

Our analysis of PIP4K/PIP5K protein sequences has identified several useful CSIs, which provide important insights into the evolution of this family of proteins. Of these CSIs, two CSIs are shared by all PIP4K homologs, but they are lacking in all PIP5K homologs as well as the single homolog of the PIP4K/PIP5K found in the deeper branching eukaryotic organisms and plants and fungi (Figures 2 and 3). The first of these CSIs consists of 1 aa insert in the kinase homology domain (Figure 2), whereas the second CSI is a 2 aa deletion in the N-terminal domain of the PIP4K protein (Figure 3). Sequence information for these CSIs for PIP4K/PIP5K homologs for representative species from the major groups of eukaryotes are shown in Figures 2 and 3. More detailed information regarding the species distribution of these CSIs is provided in Figures S2 and S3.

Sequence alignments of the PIP4K/PIP5K proteins shown in Figures 2 and 3 illustrate that the distinct homologs of both PIP4K and PIP5K are present in different Deutrostomia and Protostomia species as well in other deeper branching phyla of Animalia such as Placozoa (*Trichoplax adhaerens*), Porifera (*Amphimedon queenslandica*), and Cnidaria (*Exaiptasia pallidia* and *Hydra vulgaris*). Further, in addition to the Animalia species, distinct homologs of PIP4K and PIP5K are also present in the Choanoflagellates (*Salpinogoeca rosetta* and *Monosiga brevicollis*) and Filasterea (*C. owczarzaki*) species whose members are indicated to be the closest unicellular relatives of multicellular Animalia [57,58]. As seen from Figures 2 and 3, the two CSIs in the PIP4K homologs are shared by all orthologs of this protein including those from the unicellular metazoan phyla viz. Choanoflagellate and Filasterea, but they are not present in any of the PIP5K homologs from corresponding phyla. Further, the single homolog of PIP4K/PIP5K proteins found in the deeper branching eukaryotic lineages, such as Apicompelxa, Amebozoa, Percolozoa, and Apusozoa also lack the indicated CSIs. The absence of these CSIs in the deeper branching eukaryotic phyla indicates that the 1 aa CSI constitutes an insert in the PIP4K family of proteins, whereas the 2 aa CSI is a deletion in this protein. Based on the unique shared presence of these CSIs in all PIP4K homologs, the genetic changes responsible for these CSIs are postulated to have occurred in a common ancestor of the PIP4K family of protein at the time when the PIP4K and PIP5K family of proteins diverged from a common ancestor by a gene duplication event.

Another important CSI identified by our analysis consists of a 1aa deletion within the highly conserved kinase homology domain that is commonly shared by all PIP4K and PIP5K homologs (and their distinct isoforms), but which is not found in the single homolog of these proteins present in fungi, plants, and the deep branching eukaryotic lineages (Figure 4).

The shared presence of this CSI in different PIP4K and PIP5K homologs from all metazoan species well as in the Choanoflagellates and Filasterea phyla, which together comprise the Holozoa clade of organisms [59,60], strongly suggests that the genetic change responsible for this CSI occurred in a common ancestor of the Holozoa. Further, due to the shared presence of this CSI in both PIP4K and PIP5K homologs, the genetic change leading to this CSIs is postulated to have occurred before gene duplication leading to distinct forms of the PIP4K and PIP5K proteins, which are present in different Holozoa species.

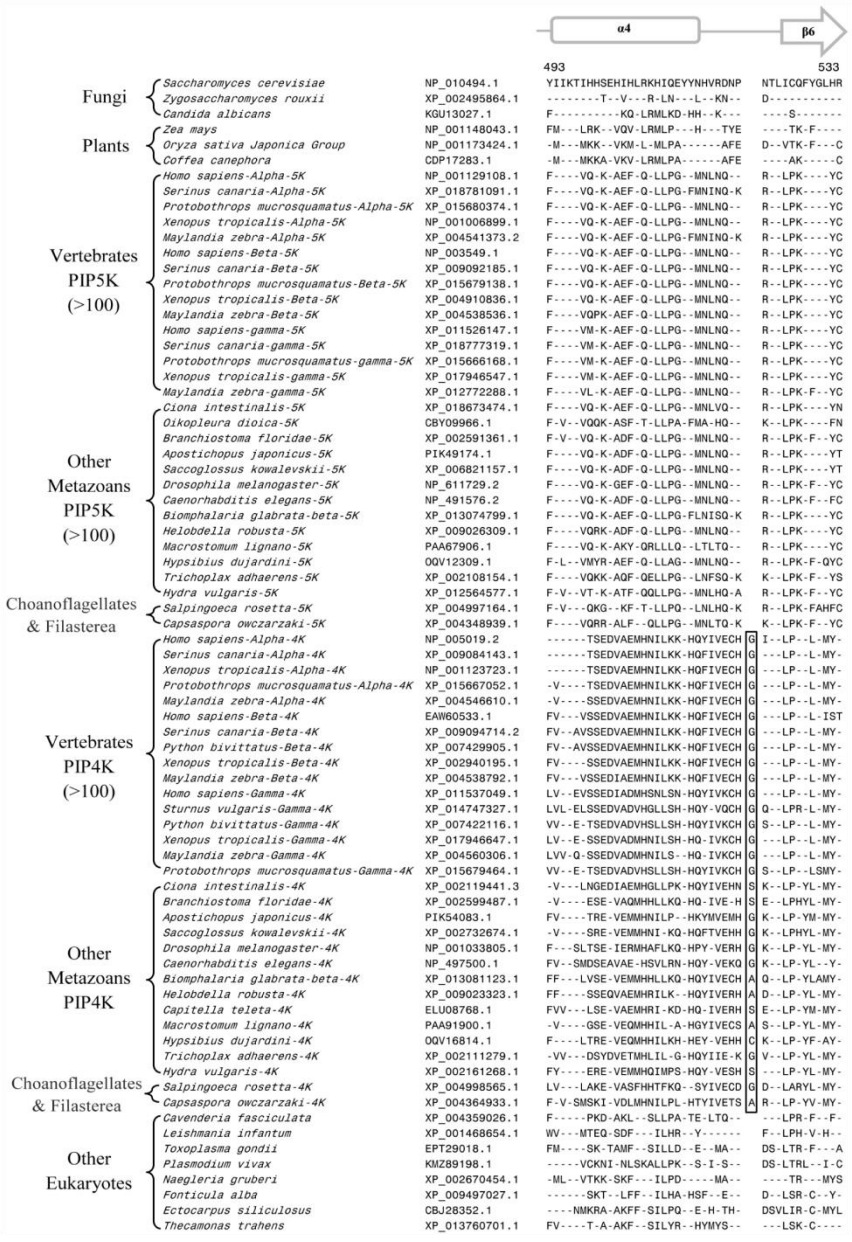


Figure 2. Excerpts from the sequence alignment of PIP4K and PIP5K homologs showing a 1 aa insert (boxed) in a conserved region that is uniquely shared by all PIP4K homologs. This insert is commonly shared by all PIP4K homologs from metazoan phyla including the Choanoflagellates and Filasterea but it is not found in any PIP5K homologs or the single orthologs of PIP4K/PIP5K found in the deeper-branching eukaryotic organisms. Detailed information regarding the species distribution of this conserved signature indels provided in Figure S2. The dashes (-) in the alignment indicates identity with the amino acids on the top line. Numbers on the top indicate the location of this sequence region in the protein from *Saccharomyces cerevisiae*. The accession numbers of various sequences are given in the second column. Secondary structure information for this protein region is presented on top of the sequence.

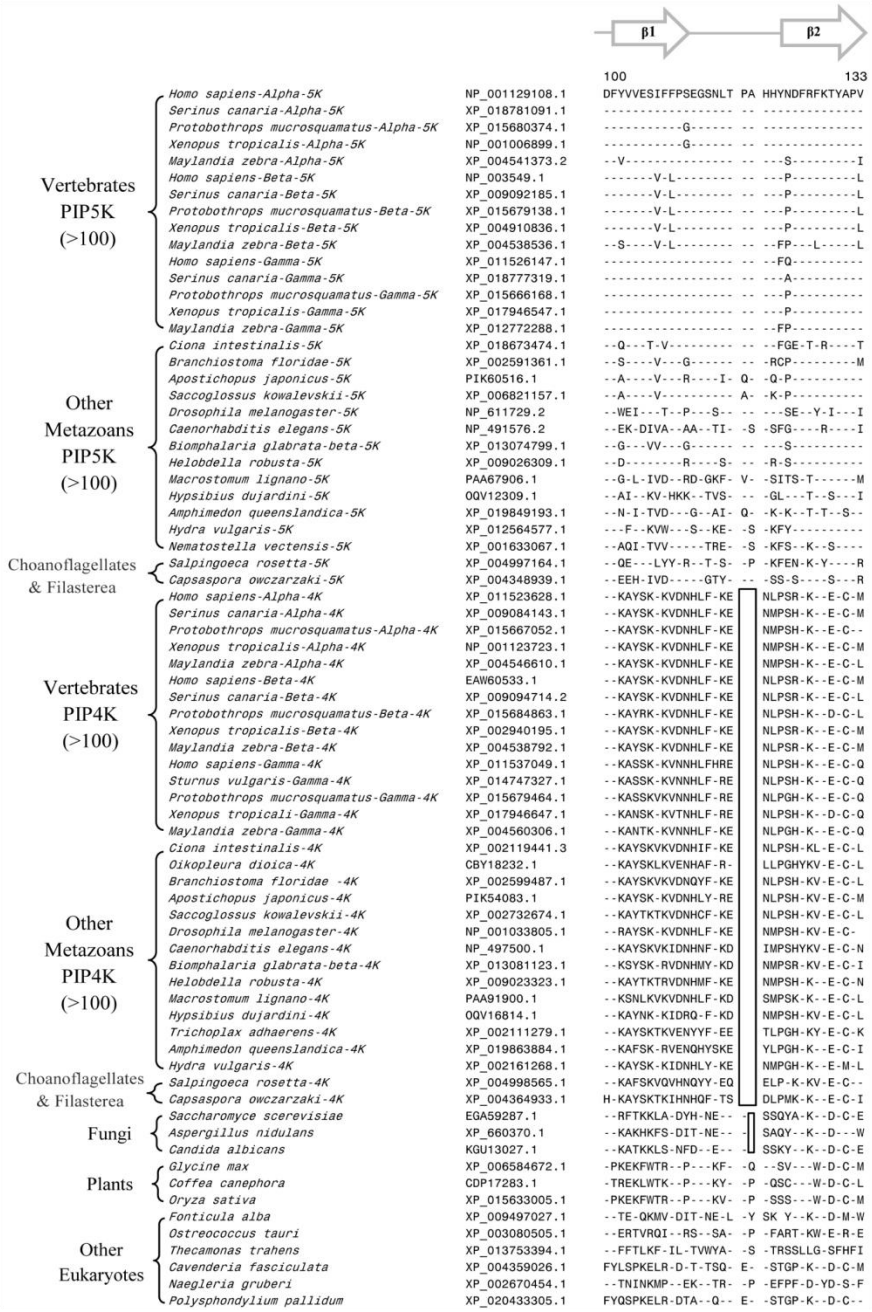


Figure 3. Partial sequence alignment of the PIP4K/PIP5K family of proteins showing a 2 aa deletion in a conserved region (boxed) that is uniquely shared by all PIP4K homologs. The boxed CSI is not present in any of the PIP5K homologs as well as the PIP4K/PIP5K orthologs from plants and other deep-branching eukaryotic lineages. The PIP4K/PIP5Korthologs from fungi contain a shorter 1 aa deletion in this position. More detailed sequence information for this CSI is provided in Figure S3. Other details are the same as in Figure 2 legend. Numbers on the top indicate the position of this sequence in the human PIP5Kα.

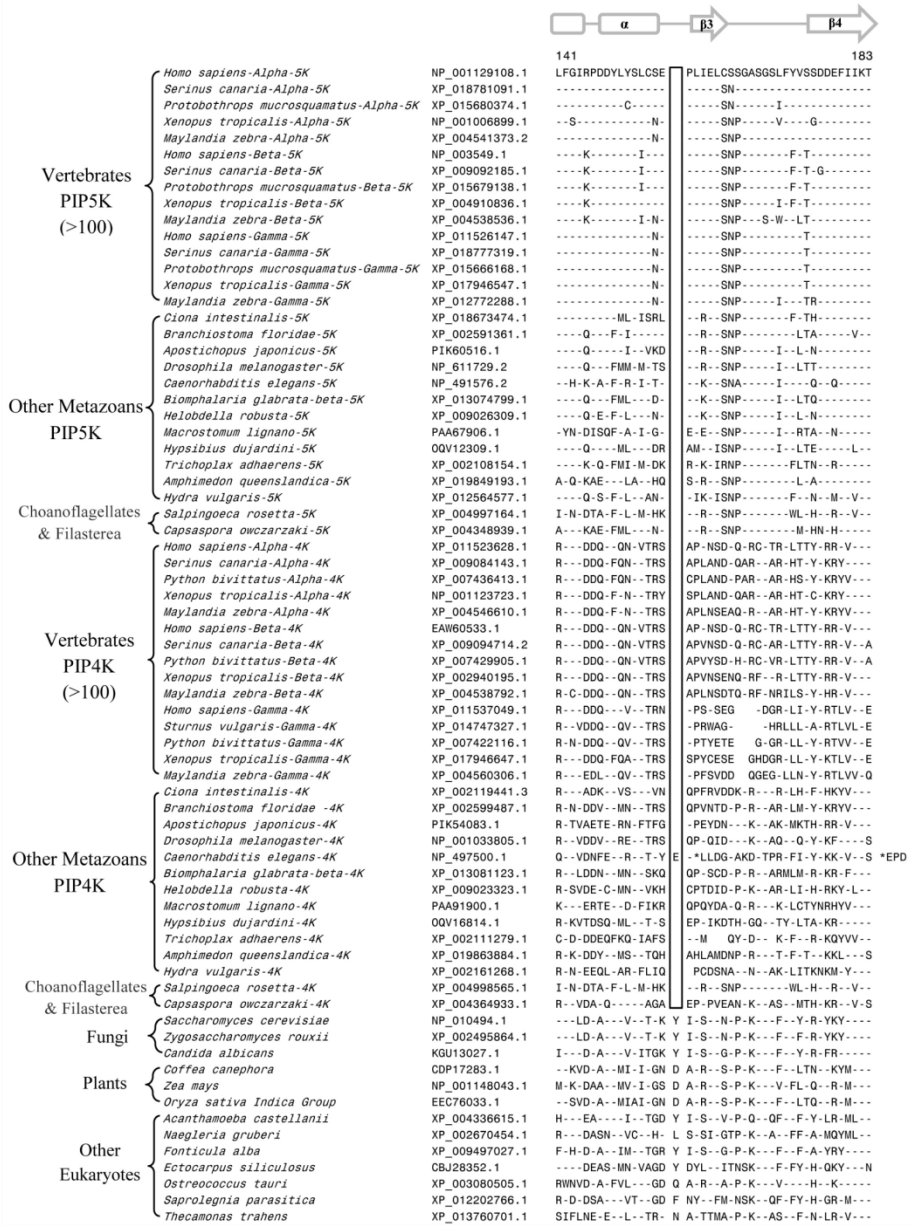


Figure 4. Partial sequence alignment of PIP4K/PIP5K family of proteins showing 1 aa deletion (boxed) in a conserved region that is commonly shared by different PIP4K and PIP5K homologs but lacking in the single copy PIP4K/PIP5K orthologs from deeper branching eukaryotic phyla including plants and fungi. This 1 aa CSI distinguishes the distinct PIP4K and PIP5K homologs found in Holozoa species (i.e., all multicellular metazoan species and their unicellular relatives Choanoflagellates and Filasterea) from early branching eukaryotic organisms harboring only a single ortholog of the PIP4K/PIP5K protein. The PIP4K from some nematode species lack this deletion or contain a longer insertion (indicated by *) in this position and its possible significance is unclear. Numbers on the top indicate the position in human PIP5Kα. More detailed information for this CSI is provided in Figure S4.

In addition to the CSIs that are specific for the PIP4K or PIP5K proteins, or both, our analysis has identified three other CSIs that are uniquely shared by members of specific subfamilies of the PIP4K and PIP5K proteins, distinguishing them from other related proteins and providing useful insights concerning their evolution. The first of these subfamily-specific CSIs consists of a 1 aa deletion that is uniquely present in all PIP5Kα homologs but not found in the PIP5Kβ and PIP5Kγ homologs (Figure 5). The observed CSI (1aa deletion) is located within the C-terminal region of the conserved core kinase homology domain, as shown in Figure 5. The region where this CSI is found is conserved in the PIP5K family of proteins, but it is lacking in the PIP4K homologs. Due to the specificity of this CSI for PIP5Kα, the genetic change responsible for this CSI is postulated to have occurred in a common ancestor of the PIP5Kα proteins, and it provides a reliable molecular characteristic distinguishing the PIP5Kα from PIP5Kβ and PIP5Kγ subfamily of proteins.

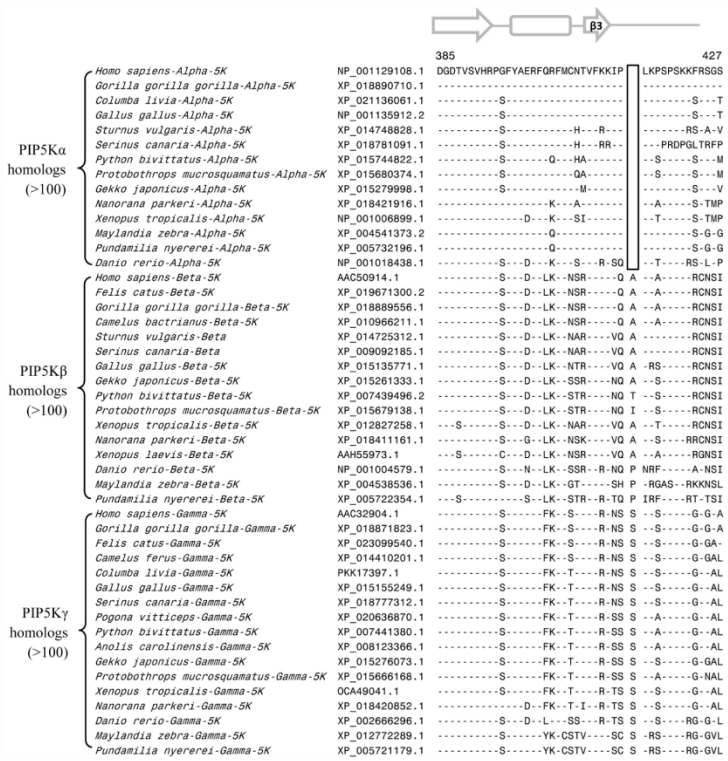


Figure 5. Partial sequence alignment of different subfamilies of the PIP5K protein showing 1 aa CSI (boxed) that is uniquely shared by the PIP5Kα subfamily of proteins. More detailed information regarding species distribution of this CSI is provided in Fig. S5. The predicted secondary structure of this sequence region is shown on top of the sequence alignment. Other details are the same as in the Figure 2 legend.

Another subfamily-specific CSI identified in our work consists of a 2 aa conserved insert that is commonly shared by the PIP5Kβ family of proteins from mammals, birds, and reptile species, but not found in the PIP5Kβ homologs from amphibians and fish (Figure 6). This CSI is absent in the PIP5Kα and PIP5Kγ family members and also in different PIP5K homologs. Based on its distribution pattern, the genetic change leading to this CSI is postulated to have occurred in a common ancestor of the

PIP5Kβ from mammals, birds, and reptiles and it supports the latter divergence of these vertebrates classes in comparison to the fishes and amphibians [61].

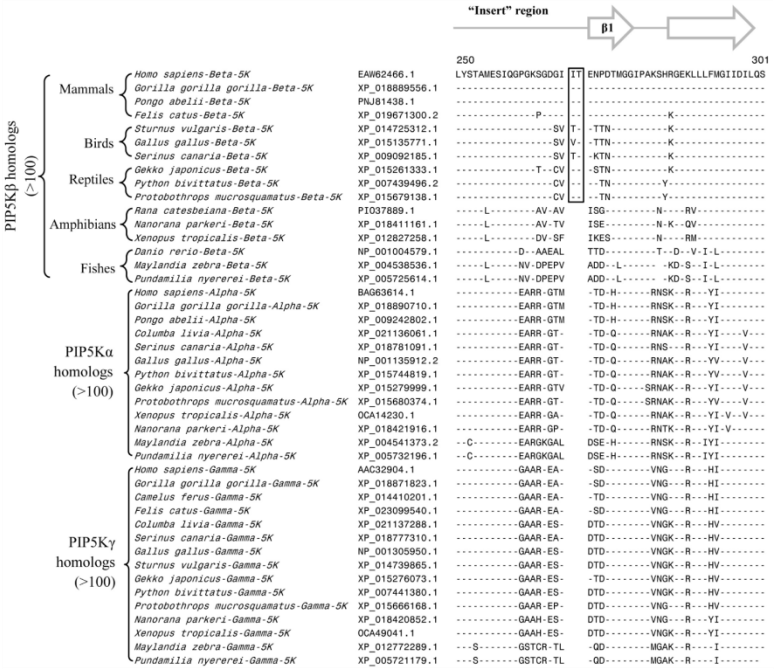


Figure 6. Excerpts from the multiple sequence alignment of PIP5K homologs showing a 2 aa insert in a conserved region (boxed) that is uniquely shared by the PIP5Kβ homologs of mammals, birds, and reptiles, but absent from all other PIP4K and PIP5K homologs. Sequence information is shown here for only representative species from different vertebrate classes. More detailed information for this CSI is provided in Figure S6. Other details are as in Figure 2 legend. Numbers on the top indicate the position in the human PIP5Kβ.

The last of the subfamily-specific CSIs is present in a highly conserved region of the PIP4Kγ, where deletions ranging from 1 aa to 4 aa are present in different groups/classes of the vertebrates (Figure 7). As seen from Figure 7, all PIP4Kγ homologs from mammals and reptiles contain a 3 aa deletion in this region. In the same place, the available homologs of PIP4Kγ from birds harbor a 4 aa deletion while those from the amphibians and fish contain shorter deletions (viz. 1-2 aa) (Figure 7). In contrast, no deletion was present in this region in the PIP4K α and β homologs from vertebrates as well as the PIP4K homologs from other metazoan species. The genetic characteristics of this CSI suggest that the observed variation in the length of this CSI in the vertebrate classes is likely the result of successive genetic changes occurring within the same region at different stages in the evolution of vertebrates. However, the possibility that the observed differences in the lengths of this CSI in the vertebrate lineages are due to independent genetic changes cannot be ruled out. Nonetheless, the observance of these lineage-specific genetic changes within this conserved region of PIP4Kγ suggests that this region plays an important role in determining functional characteristics which are specific for the different classes of vertebrates. Indeed, the functional significance of this region has been reported in a number of recent studies [11,43]. Clarke and Irvine [11] have examined the functional significance of this region with respect to the catalytic activities and substrate interactions of PIP4K isoforms by creating a series of mutations in the PIP4Kγ by replacing the residues of G-loop from the PIP4Kα sequence.

The mutant forms of PIP4Kγ containing these changes showed a significant increase in lipid kinase activity [11], suggesting that the evolutionarily conserved indels in this region play an important role in determining the functional differences between different isoforms of PIP4K.

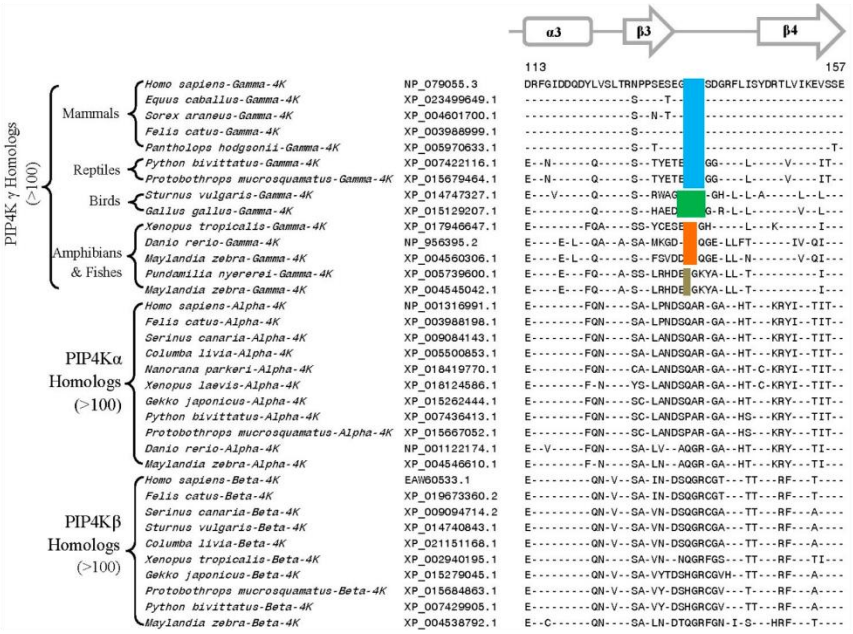


Figure 7. Partial sequence alignment of different isoforms of the PIP4K proteins showing a conserved region where 1–4 aa deletions (boxed) are uniquely found in the PIP4Kγ subfamily of proteins from different classes of vertebrates. The PIP4Kγ protein mammals and reptiles have a 3 aa deletion in this position, whereas those from the birds contain a 4 aa deletion in the same position. The fish and amphibians are found to contain multiple copies of PIP4Kγ with 1–2 aa deletion in this position. More detailed information for this CSI is provided in Figure S7. Other details are the same as in Figure 2 legend. Numbers on the top indicate the position of the sequence in the human PIP4Kγ.

3.3. Locations of the Identified CSIs in the Structures of the PIP4K/PIP5K Proteins

Earlier work on CSIs in different proteins show that most, if not all, of the previously studied CSIs are located on the surface exposed loops of different proteins, which are important in mediating novel protein-protein or protein-ligands interaction [38,62–65]. An 8 aa CSI in the PIP5K protein identified in our earlier work, which was specific for the *Saccharomycetaceae* family of fungi, was also located in a surface exposed loop and it was indicated to play an important role in the binding of this protein with membrane surface [33]. In view of these earlier studies, it was of interest to determine the locations in protein structures of the different CSIs identified in the present work. For these studies, structural information that was currently available for different solved structures for the PIP4K/PIP5K family of proteins listed in the Methods section was utilized. In addition, a homology model was also created for the human PIP5Kβ using the solved structure of PIP5Kα from *Danio rerio* (PDB ID: 4TZ7, chain "A") as a template. In Figure 8, we show a composite diagram, wherein we have mapped the locations of most of the identified CSIs in a structural model of the human PIP5Kβ protein. The CSIs which represent inserts are marked in red whereas CSIs representing deletions are shown in blue. In the zoomed regions in this figure, the structural models showing the locations of different CSIs are shown in cartoon representation. There was no structural information available for the region where the CSI

shown in Figure 5 in PIP5Kα is found. Hence, its location in the protein structure was not mapped. As shown in this figure, all identified CSIs in the PIP4K/PIP5K family of proteins are found to be located on the surface exposed loop region and thus they should be able to interact with other proteins/ligands.

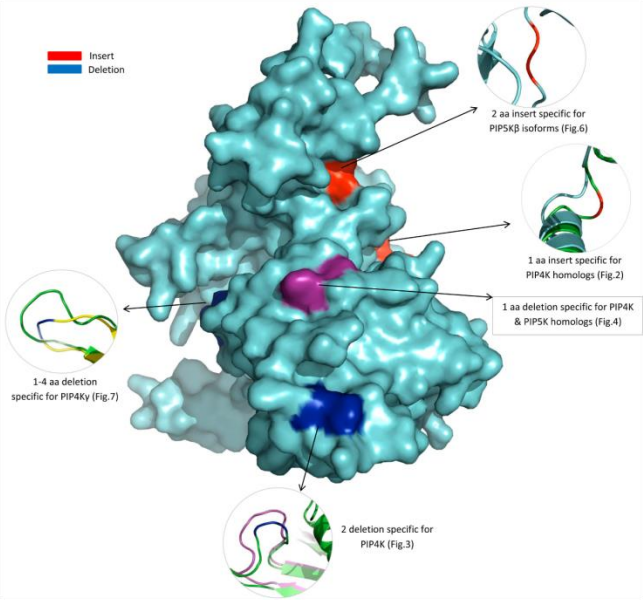


Figure 8. Surface representation of the identified CSIs in a structural model of the human PIP5Kβ protein. For mapping of the CSIs in protein structures, structural information for a number of solved/modeled structures for the PIP4K/PIP5K family of proteins (see Methods section) was utilized. The CSIs which constitute inserts are marked in red on the surface, while for the CSIs that are deletions, the protein regions where these deletions are found are marked in blue on the surface. The location of the 1 aa deletion (Figure 4) that is commonly shared by different homologs of PIP4K and PIP5K is shown in magenta based on structural comparison with the *Saccharomyces cerevisiae* PIP4K/PIP5K homolog. The close-up views of the locations in the protein structure for different identified CSIs are shown in cartoon representation. The structure model of PIP4Kγ isoform from *H. sapiens* is shown in yellow and crystal structure of PIP4Kβ isoform is shown in green.

4. Discussion

The PIP4K/PIP5K family of proteins constitutes crucial players in the regulation of the metabolism of phosphatidylinositides in eukaryotes [21–23]. Both PIP4K and PIP5K are involved in generating a key signaling molecule, PI(4,5)P_2, which resides at the core of the phosphatidylinositol signaling pathway controlling a wide range of fundamental cellular processes [1,3,5]. Further, due to the important roles played by these proteins in many critical processes involved in pathological conditions [66–71] these proteins are becoming an increasingly interesting class of molecular targets for cancer [72,73], chronic pain [74], diabetes [75], and autoimmune diseases [76]. However, despite the important roles played by these proteins in the regulation of many cellular processes, our understanding of the overall evolutionary relationships between different members of the PIP4K/PIP5K families and subfamilies of proteins and what specific genetic/biochemical characteristics distinguish different members of this protein family remains limited.

The present study was undertaken with the aims of advancing our understanding of the evolutionary divergence of different members of the PIP4K/PIP5K protein families from a common

ancestor and also to identify novel molecular features in these proteins that are distinctive characteristics of a particular family or subfamily of these proteins. Our analyses of protein sequences from the PIP4K/PIP5K reported here have identified six highly-specific molecular signatures in the forms of CSIs that are distinctive characteristics of specific isozymes and subfamilies of these proteins. Based on the results obtained from the species distributions of PIP4K/PIP5K isozymes/homologs in different eukaryotic lineages, phylogenetic analysis based on the sequences of these proteins, and the inferences derived from the species and isozyme specificities of different identified CSIs in these proteins, several novel and important insights regarding the evolutionary divergence of this protein family can be gleaned. An overall summary of the results obtained from different approaches regarding the evolutionary divergence of the PIP4K/PIP5K family of proteins is presented in Figure 9. In addition to showing the distribution of different members of the PIP4K/PIP5K family of proteins in eukaryotic lineages, this figure also marks the evolutionary stages where the genetic changes responsible for different identified CSIs likely occurred during the evolution of this protein family.

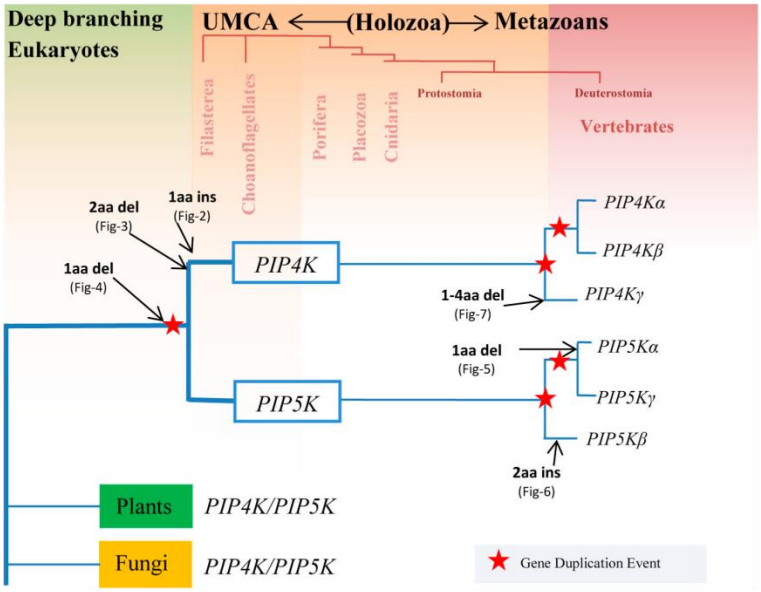


Figure 9. A summary diagram showing the evolutionary divergence of different members of the PIP4K/PIP5K family of proteins in eukaryotic organisms. The model presented here is based on the species distribution of different proteins as well as the species/isozyme specificities of different CSIs in these proteins that were identified in the present work. The arrows mark the evolutionary stages where the rare genetic changes leading to the specific CSI(s) are postulated to have occurred. The red stars mark the evolutionary stages where gene duplication events have occurred during the divergence of this protein family. Holozoa clade comprises of the multicellular metazoan phyla as well as phyla consisting of their unicellular metazoan common ancestor (UMCA).

As shown here, the distinct homologs of both PIP4K and PIP5K, in addition to the Bilateria species [4], are also found in other multicellular metazoan phyla (viz Cnidaria, Placozoa, and Porifera) as well as in the Choanoflagellates and Filasterea, which are the closest-known unicellular ancestor of the multicellular animals. These groups together form the Holozoa clade of species [59,60]. In contrast, all other deeper branching eukaryotic lineages including Apicompelxa, Amebozoa, Percolzoa, and Apusozoa, as well as plants and fungi, contain only a single homolog of these proteins showing similarity to both PIP4K/PIP5K. Our work has identified a conserved indel that is commonly shared by

all PIP4K and PIP5K homologs from Holozoa species (except some PIP4K homologs from nematodes, whose significance is unclear) but absent in the single copy homologs of these proteins found in the deeper branching lineages (Figure 4). Based on the species and isozyme distribution of this CSI, the genetic change leading to this CSI has occurred in a common ancestor of the Holozoa clade prior to the first gene duplication event leading to the divergence of distinct types of PIP5K and PIP4K homologs. These results indicate that this genetic change preceded the evolution of all animals (i.e., both unicellular and multicellular). Two other CSIs identified in this study are uniquely found in all PIP4K homologs (Figures 2 and 3) but they not present in any PIP5K homolog. The genetic changes giving rise to these CSIs are postulated to have occurred in a common ancestor of the PIP4K homologs soon after the gene duplication event leading to the formation of distinct forms of the PIP4K and PIP5K proteins. Due to the specificities of these CSIs for either the PIP4K/PIP5K proteins or only the PIP4K isozymes, these molecular markers provide useful means for genetic and biochemical studies leading to the discovery of novel and distinctive properties of these isozymes.

The vertebrate species contain three different subfamilies of PIP4K and PIP5K proteins. In phylogenetic trees, all three subfamilies of the PIP4K are part of one distinct clade whereas the different members of the PIP5K subfamilies form a separate clade. As noted in earlier work [4], the branching patterns of these proteins strongly suggest that the members of these subfamilies have originated from two independent and successive duplications of the genes for PIP4K and PIP5K proteins within vertebrates [4]. As all three forms (or isoforms) of PIP4K and PIP5K proteins are present in different (or most) vertebrate species, the gene duplication events leading to the divergence of these protein families have occurred in a vertebrates' common ancestor within an evolutionarily short period. Due to this, the relative branching orders of these protein families can be inferred only tentatively. Work on the evolution of the phosphatidylinositol-3-kinases family of protein kinases indicates that the gene for the catalytic subunit of this protein has also undergone two major duplication events at different stages in the evolution of eukaryotic organisms to account for its species distribution [77].

It is important to understand how different isoforms of the PIP4K and PIP5K proteins differ from each other and what unique features distinguish these isoforms. In this context, our identification of several CSIs that are distinctive characteristics of the specific subfamilies of either PIP4K or PIP5K proteins, or both, is of much interest. Of the three subfamily-specific CSIs identified in this work, one is specific for all PIP5Kα homologs. The genetic change responsible for this CSI likely occurred in a common ancestor of the PIP5Kα subfamily, when it diverged from the PIP5Kβ-PIP5Kγ by a gene duplication event. Another CSI described here is specific for the PIP5Kβ homologs from mammals, birds, and reptiles (Figure 6). The genetic change leading to this CSI likely occurred in a common ancestor of the PIP5Kβ protein from mammals, birds, and reptiles and it supports the latter divergence of these vertebrate classes in comparison to the fish and amphibians [61]. In addition to these CSIs, within a highly conserved region of the PIP4Kγ subfamily of proteins, deletions of specific lengths are present in different groups/classes of vertebrate species (Figure 7). The observed variation in the length of this CSI in vertebrate groups suggests that successive genetic changes have occurred in this position during the evolution of vertebrates, indicating that this region should be of particular importance in the functioning of this protein. The genetic changes leading to these isoforms or subfamilies-specific CSIs have occurred at important stages in the evolution of vertebrate species and the identified molecular signatures provide important means for distinguishing some of these distinct isoforms and understanding their unique functional characteristics.

As has been noted earlier, the CSIs in genes/proteins sequences result from rare genetic changes and these changes have been found to be important/essential for the proper functioning of the proteins in the CSI-containing organisms [33,38,78]. Further, most studied CSIs, including those identified in this study, are located in surface-exposed loops of the proteins which, due to their ability to interact with other proteins and ligands, perform important roles in mediating novel functional interactions [33,38,62–64,79]. Previously, we have reported the identification of an 8 aa CSI in the PIP5K homologs which was specific for the *Saccharomycetaceae* family of fungi. This CSI formed a

positively-charged patch on the surface of the protein and it was predicted to play a role in the binding of the yeast PIP5K protein with the membrane. This prediction was strongly supported by molecular dynamics simulation studies examining the binding interaction of the yeast PIP5K protein, with and without the CSI, with the membrane lipid bilayers. Clarke and Irvine [11] have examined the functional significance of the region where the 1–4 aa deletions are present in the PIP4Kγ isoforms in different vertebrate species. Their study also showed that the changes in the residues corresponding to the identified CSIs significantly affected the functional activity of the protein and could account for the differences in the functional activity of the different isoforms [11]. Thus, it is strongly expected that the other CSIs identified in this study in the PIP4K/PIP5K family of proteins should also be playing important roles in the cellular functions of these proteins and the described CSIs provide novel genetic and biochemical means to investigate such differences. Lastly, due to the important roles played by the PIP4K/PIP5K family of proteins in different cellular processes associated with a variety of pathological conditions (e.g., cancer, diabetes, chronic pain, autoimmune diseases) [66–71], the identified CSIs in these proteins, which are surface exposed and predicted to play important cellular functions, also provide potential means for development of novel therapeutics targeting specific diseases [72–76].

Supplementary Materials: The following are available online at http://www.mdpi.com/2073-4425/10/4/312/s1, Table S1: Sequence information for different PIP4K/PIP5K family of protein sequences used in phylogenetic studies, Figure S2: detailed species distribution information for the 1 aa CSI in PIP4K shown in Figure 2, Figure S3: detailed species distribution information for the 2 aa deletion in PIP4K in Figure 3, Figure S4:; detailed species distribution information for the 1 aa CSI shown in Figure 4, Figure S5: detailed species distribution information for the 1 aa conserved deletion in PIP5Kα shown in Figure 5, Figure S6: detailed species distribution information for the 2 aa conserved insert in PIP5Kβ isoform shown in Figure 6, Figure S7: detailed species distribution information for the 1–4 aa conserved deletions in PIP4Kγ isoforms shown in Figure 7.

Author Contributions: B.K.: Identification of CSIs and confirming their species/homolog specificities, species distribution of PIP4K/PIP5K homologs, phylogenetic analysis, homology modelling and localization of the CSIs in protein structures, preparation of draft manuscript; R.S.G., planning and supervision of the entire work, interpretation of the results and writing and finalizing of the manuscript, obtained funding for the project.

Funding: This work was supported by Research Grant number 249924 from the Natural Science and Engineering Research Council of Canada awarded to Radhey S. Gupta.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Di Paolo, G.; De Camilli, P. Phosphoinositides in cell regulation and membrane dynamics. *Nature* **2006**, *443*, 651–657. [CrossRef]
2. Balla, T.; Szentpetery, Z.; Kim, Y.J. Phosphoinositide signaling: New tools and insights. *Physiology* **2009**, *24*, 231–244. [CrossRef]
3. Balla, T. Phosphoinositides: Tiny lipids with giant impact on cell regulation. *Physiol. Rev.* **2013**, *93*, 1019–1137. [CrossRef]
4. Brown, J.R.; Auger, K.R. Phylogenomics of phosphoinositide lipid kinases: Perspectives on the evolution of second messenger signaling and drug discovery. *BMC. Evol. Biol.* **2011**, *11*, 4. [CrossRef]
5. Loijens, J.C.; Boronenkov, I.V.; Parker, G.J.; Anderson, R.A. The phosphatidylinositol 4-phosphate 5-kinase family. *Adv. Enzyme Regul.* **1996**, *36*, 115–140. [CrossRef]
6. Aikawa, Y.; Martin, T.F. ARF6 regulates a plasma membrane pool of phosphatidylinositol(4,5)bisphosphate required for regulated exocytosis. *J. Cell. Biol.* **2003**, *162*, 647–659. [CrossRef]
7. Mellman, D.L.; Gonzales, M.L.; Song, C.; Barlow, C.A.; Wang, P.; Kendziorski, C.; Anderson, R.A. A PtdIns4,5P_2-regulated nuclear poly(A) polymerase controls expression of select mRNAs. *Nature* **2008**, *451*, 1013–1017. [CrossRef] [PubMed]
8. Ishihara, H.; Shibasaki, Y.; Kizuki, N.; Wada, T.; Yazaki, Y.; Asano, T.; Oka, Y. Type I phosphatidylinositol-4-phosphate 5-kinases. Cloning of the third isoform and deletion/substitution analysis of members of this novel lipid kinase family. *J. Biol. Chem.* **1998**, *273*, 8741–8748.

9. Ishihara, H.; Shibasaki, Y.; Kizuki, N.; Katagiri, H.; Yazaki, Y.; Asano, T.; Oka, Y. Cloning of cDNAs encoding two isoforms of 68-kDa type I phosphatidylinositol-4-phosphate 5-kinase. *J. Biol. Chem.* **1996**, *271*, 23611–23614. [CrossRef] [PubMed]
10. Clarke, J.H.; Emson, P.C.; Irvine, R.F. Localization of phosphatidylinositol phosphate kinase IIgamma in kidney to a membrane trafficking compartment within specialized cells of the nephron. *Am. J. Physiol. Renal. Physiol.* **2008**, *295*, F1422–F1430. [CrossRef]
11. Clarke, J.H.; Irvine, R.F. Evolutionarily conserved structural changes in phosphatidylinositol 5-phosphate 4-kinase (PI5P4K) isoforms are responsible for differences in enzyme activity and localization. *Biochem. J.* **2013**, *454*, 49–57. [CrossRef] [PubMed]
12. Xia, Y.; Irvine, R.F.; Giudici, M.L. Phosphatidylinositol 4-phosphate 5-kinase Igamma_v6, a new splice variant found in rodents and humans. *Biochem. Biophys. Res. Commun.* **2011**, *411*, 416–420. [CrossRef] [PubMed]
13. Shulga, Y.V.; Anderson, R.A.; Topham, M.K.; Epand, R.M. Phosphatidylinositol-4-phosphate 5-kinase isoforms exhibit acyl chain selectivity for both substrate and lipid activator. *J. Biol. Chem.* **2012**, *287*, 35953–35963. [CrossRef] [PubMed]
14. Desrivieres, S.; Cooke, F.T.; Parker, P.J.; Hall, M.N. MSS4, a phosphatidylinositol-4-phosphate 5-kinase required for organization of the actin cytoskeleton in Saccharomyces cerevisiae. *J. Biol. Chem.* **1998**, *273*, 15787–15793. [CrossRef] [PubMed]
15. Guillas, I.; Vernay, A.; Vitagliano, J.J.; Arkowitz, R.A. Phosphatidylinositol 4,5-bisphosphate is required for invasive growth in Saccharomyces cerevisiae. *J. Cell Sci.* **2013**, *126*, 3602–3614. [CrossRef] [PubMed]
16. Okazaki, K.; Miyagishima, S.Y.; Wada, H. Phosphatidylinositol 4-phosphate negatively regulates chloroplast division in Arabidopsis. *Plant Cell* **2015**, *27*, 663–674. [CrossRef]
17. Heilmann, I. Phosphoinositide signaling in plant development. *Development* **2016**, *143*, 2044–2055. [CrossRef]
18. Mueller-Roeber, B.; Pical, C. Inositol phospholipid metabolism in Arabidopsis. Characterized and putative isoforms of inositol phospholipid kinase and phosphoinositide-specific phospholipase C. *Plant Physiol.* **2002**, *130*, 22–46. [CrossRef] [PubMed]
19. Xue, H.W.; Chen, X.; Mei, Y. Function and regulation of phospholipid signalling in plants. *Biochem. J.* **2009**, *421*, 145–156. [CrossRef]
20. Shisheva, A. PIKfyve: Partners, significance, debates and paradoxes. *Cell Biol. Int.* **2008**, *32*, 591–604. [CrossRef]
21. Oude Weernink, P.A.; Schmidt, M.; Jakobs, K.H. Regulation and cellular roles of phosphoinositide 5-kinases. *Eur. J. Pharmacol.* **2004**, *500*, 87–99. [CrossRef]
22. Van den Bout, I.; Divecha, N. PIP5K-driven PtdIns(4,5)P2 synthesis: Regulation and cellular functions. *J. Cell Sci.* **2009**, *122*, 3837–3850. [CrossRef]
23. Bulley, S.J.; Clarke, J.H.; Droubi, A.; Giudici, M.L.; Irvine, R.F. Exploring phosphatidylinositol 5-phosphate 4-kinase function. *Adv. Biol. Regul.* **2015**, *57*, 193–202. [CrossRef]
24. Baldauf, S.L.; Palmer, J.D. Animals and fungi are each other's closest relatives: Congruent evidence from multiple proteins. *Proc. Natl. Acad. Sci. USA* **1993**, *90*, 11558–11562. [CrossRef]
25. Gupta, R.S.; Aitken, K.; Falah, M.; Singh, B. Cloning of Giardia lamblia heat shock protein HSP70 homologs: Implications regarding origin of eukaryotic cells and of endoplasmic reticulum. *Proc. Natl. Acad. Sci. USA* **1994**, *91*, 2895–2899. [CrossRef]
26. Gupta, R.S. Protein phylogenies and signature sequences: A reappraisal of evolutionary relationships among archaebacteria, eubacteria, and eukaryotes. *Microbiol. Mol. Biol. Rev.* **1998**, *62*, 1435–1491. [PubMed]
27. Gupta, R.S. Molecular signatures that are distinctive characteristics of the vertebrates and chordates and supporting a grouping of vertebrates with the tunicates. *Mol. Phylogenet. Evol.* **2016**, *94*, 383–391. [CrossRef]
28. Rivera, M.C.; Lake, J.A. Evidence that eukaryotes and eocyte prokaryotes are immediate relatives. *Science* **1992**, *257*, 74–76. [CrossRef]
29. Rokas, A.; Holland, P.W. Rare genomic changes as a tool for phylogenetics. *Trends Ecol. Evol.* **2000**, *15*, 454–459. [CrossRef]
30. Gupta, R.S.; Epand, R.M. Phylogenetic analysis of the diacylglycerol kinase family of proteins and identification of multiple highly-specific conserved inserts and deletions within the catalytic domain that are distinctive characteristics of different classes of DGK homologs. *PLoS ONE* **2017**, *12*, e0182758. [CrossRef] [PubMed]

31. NCBI. Database Resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* **2017**, *45*, D12–D17. [CrossRef]
32. *Methods in Microbiology New Approaches to Prokaryotics Systematics*; Goodfellow, I.C.S.M., Chun, J., Eds.; Academic Press: London, UK, 2014; pp. 153–182.
33. Khadka, B.; Gupta, R.S. Identification of a conserved 8 aa insert in the PIP5K protein in the Saccharomycetaceae family of fungi and the molecular dynamics simulations and structural analysis to investigate its potential functional role. *Proteins* **2017**, *85*, 1454–1467. [CrossRef]
34. Larkin, M.A.; Blackshields, G.; Brown, N.P.; Chenna, R.; McGettigan, P.A. Clustal W and Clustal X version 2.0. *Bioinformatics* **2007**, *23*, 2947–2948. [CrossRef]
35. Gupta, R.S. Impact of genomics on the understanding of microbial evolution and classification: The importance of Darwin's views on classification. *FEMS Microbiol. Rev.* **2016**, *40*, 520–553. [CrossRef]
36. Tamura, K.; Stecher, G.; Peterson, D.; Filipski, A.; Kumar, S. MEGA6: Molecular Evolutionary Genetics Analysis version 6.0. *Mol. Biol. Evol.* **2013**, *30*, 2725–2729. [CrossRef]
37. Jones, D.T.; Taylor, W.R.; Thornton, J.M. The rapid generation of mutation data matrices from protein sequences. *Comput. Appl. Biosci.* **1992**, *8*, 275–282. [CrossRef]
38. Gupta, R.S.; Nanda, A.; Khadka, B. Novel molecular, structural and evolutionary characteristics of the phosphoketolases from bifidobacteria and Coriobacteriales. *PLoS ONE* **2017**, *12*, e0172176. [CrossRef]
39. Khadka, B.; Adeolu, M.; Blankenship, R.E.; Gupta, R.S. Novel insights into the origin and diversification of photosynthesis based on analyses of conserved indels in the core reaction center proteins. *Photosynth. Res.* **2017**, *131*, 159–171. [CrossRef]
40. Altschul, S.F.; Madden, T.L.; Schaffer, A.A.; Zhang, J.; Zhang, Z.; Miller, W.; Lipman, D.J. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res.* **1997**, *25*, 3389–3402. [CrossRef]
41. Rose, P.W.; Prlic, A.; Bi, C.; Bluhm, W.F.; Christie, C.H.; Dutta, S.; Green, R.K.; Goodsell, D.S.; Westbrook, J.D.; Woo, J.; et al. The RCSB Protein Data Bank: Views of structural biology for basic and applied research and education. *Nucleic Acids Res.* **2015**, *43*, D345–D356. [CrossRef]
42. Hu, J.; Yuan, Q.; Kang, X.; Qin, Y.; Li, L.; Ha, Y.; Wu, D. Resolution of structure of PIP5K1A reveals molecular mechanism for its regulation by dimerization and dishevelled. *Nat. Commun.* **2015**, *6*, 8205. [CrossRef]
43. Rao, V.D.; Misra, S.; Boronenkov, I.V.; Anderson, R.A.; Hurley, J.H. Structure of type IIbeta phosphatidylinositol phosphate kinase: A protein kinase fold flattened for interfacial phosphorylation. *Cell* **1998**, *94*, 829–839. [CrossRef]
44. Shen, M.Y.; Sali, A. Statistical potential for assessment and prediction of protein structures. *Protein Sci.* **2006**, *15*, 2507–2524. [CrossRef]
45. Eswar, N.; Webb, B.; Marti-Renom, M.A.; Madhusudhan, M.S.; Eramian, D.; Shen, M.Y.; Pieper, U.; Sali, A. Comparative protein structure modeling using MODELLER. *Curr. Protoc. Protein Sci.* **2007**, *50*, 1–4. [CrossRef]
46. Xu, D.; Zhang, Y. Improving the physical realism and structural accuracy of protein models by a two-step atomic-level energy minimization. *Biophys. J.* **2011**, *101*, 2525–2534. [CrossRef]
47. Fiser, A.; Sali, A. ModLoop: Automated modeling of loops in protein structures. *Bioinformatics* **2003**, *19*, 2500–2501. [CrossRef]
48. Lovell, S.C.; Davis, I.W.; Arendall, W.B., III; de Bakker, P.I.; Word, J.M.; Prisant, M.G.; Richardson, J.S.; Richardson, D.C. Structure validation by Calpha geometry: Phi, psi and Cbeta deviation. *Proteins* **2003**, *50*, 437–450. [CrossRef]
49. Sippl, M.J. Recognition of errors in three-dimensional structures of proteins. *Proteins* **1993**, *17*, 355–362. [CrossRef]
50. Wiederstein, M.; Sippl, M.J. ProSA-web: Interactive web service for the recognition of errors in three-dimensional structures of proteins. *Nucleic Acids Res.* **2007**, *35*, W407–W410. [CrossRef]
51. Benkert, P.; Tosatto, S.C.; Schomburg, D. QMEAN: A comprehensive scoring function for model quality assessment. *Proteins* **2008**, *71*, 261–277. [CrossRef]
52. Mendoza, L.; Taylor, J.W.; Ajello, L. The class mesomycetozoea: A heterogeneous group of microorganisms at the animal-fungal boundary. *Annu. Rev. Microbiol.* **2002**, *56*, 315–344. [CrossRef] [PubMed]
53. Shalchian-Tabrizi, K.; Minge, M.A.; Espelund, M.; Orr, R.; Ruden, T.; Jakobsen, K.S.; Cavalier-Smith, T. Multigene phylogeny of choanozoa and the origin of animals. *PLoS ONE* **2008**, *3*, e2098. [CrossRef]

54. Torruella, G.; Derelle, R.; Paps, J.; Lang, B.F.; Roger, A.J.; Shalchian-Tabrizi, K.; Ruiz-Trillo, I. Phylogenetic relationships within the Opisthokonta based on phylogenomic analyses of conserved single-copy protein domains. *Mol. Biol. Evol.* **2012**, *29*, 531–544. [CrossRef]
55. Hehenberger, E.; Tikhonenkov, D.V.; Kolisko, M.; del Campo, J.; Esaulov, A.S.; Mylnikov, A.P.; Keeling, P.J. Novel Predators Reshape Holozoan Phylogeny and Reveal the Presence of a Two-Component Signaling System in the Ancestor of Animals. *Curr. Biol.* **2017**, *27*, 2043–2050. [CrossRef] [PubMed]
56. Springer, M.S.; Stanhope, M.J.; Madsen, O.; de Jong, W.W. Molecules consolidate the placental mammal tree. *Trends Ecol. Evol.* **2004**, *19*, 430–438. [CrossRef]
57. Carr, M.; Leadbeater, B.S.; Hassan, R.; Nelson, M.; Baldauf, S.L. Molecular phylogeny of choanoflagellates, the sister group to Metazoa. *Proc. Natl. Acad. Sci. USA* **2008**, *105*, 16641–16646. [CrossRef]
58. Ruiz-Trillo, I.; Roger, A.J.; Burger, G.; Gray, M.W.; Lang, B.F. A phylogenomic investigation into the origin of metazoa. *Mol. Biol. Evol.* **2008**, *25*, 664–672. [CrossRef]
59. Steenkamp, E.T.; Wright, J.; Baldauf, S.L. The protistan origins of animals and fungi. *Mol. Biol. Evol.* **2006**, *3*, 93–106. [CrossRef]
60. Lang, B.F.; O'Kelly, C.; Nerad, T.; Gray, M.W.; Burger, G. The closest unicellular relatives of animals. *Curr. Biol.* **2002**, *12*, 1773–1778. [CrossRef]
61. Kumar, S.; Hedges, S.B. A molecular timescale for vertebrate evolution. *Nature* **1998**, *392*, 917–920. [CrossRef]
62. Akiva, E.; Itzhaki, Z.; Margalit, H. Built-in loops allow versatility in domain-domain interactions: Lessons from self-interacting domains. *Proc. Natl. Acad. Sci. USA* **2008**, *105*, 13292–13297. [CrossRef]
63. Hashimoto, K.; Panchenko, A.R. Mechanisms of protein oligomerization, the critical role of insertions and deletions in maintaining different oligomeric states. *Proc. Natl. Acad. Sci. USA* **2010**, *107*, 20352–20357. [CrossRef] [PubMed]
64. Alnajar, S.; Khadka, B.; Gupta, R.S. Ribonucleotide Reductases from Bifidobacteria Contain Multiple Conserved Indels Distinguishing Them from All Other Organisms: In Silico Analysis of the Possible Role of a 43 aa Bifidobacteria-Specific Insert in the Class III RNR Homolog. *Front. Microbiol.* **2017**, *8*, 1409. [CrossRef]
65. Hassan, F.M.N.; Gupta, R.S. Novel Sequence Features of DNA Repair Genes/Proteins from Deinococcus Species Implicated in Protection from Oxidatively Generated Damage. *Genes* **2018**, *9*, 149. [CrossRef] [PubMed]
66. Singhal, R.L.; Prajda, N.; Yeh, Y.A.; Weber, G. 1-Phosphatidylinositol 4-phosphate 5-kinase (EC 2.7.1.68): A proliferation- and malignancy-linked signal transduction enzyme. *Cancer Res.* **1994**, *54*, 5574–5578.
67. Luoh, S.W.; Venkatesan, N.; Tripathi, R. Overexpression of the amplified Pip4k2beta gene from 17q11–12 in breast cancer cells confers proliferation advantage. *Oncogene* **2004**, *23*, 1354–1363. [CrossRef] [PubMed]
68. Schleiermacher, G.; Bourdeaut, F.; Combaret, V.; Picrron, G.; Raynal, V.; Aurias, A.; Ribeiro, A.; Janoueix-Lerosey, I.; Delattre, O. Stepwise occurrence of a complex unbalanced translocation in neuroblastoma leading to insertion of a telomere sequence and late chromosome 17q gain. *Oncogene* **2005**, *24*, 3377–3384. [CrossRef]
69. Narkis, G.; Ofir, R.; Landau, D.; Manor, E.; Volokita, M.; Hershkowitz, R.; Elbedour, K.; Birk, O.S. Lethal contractural syndrome type 3 (LCCS3) is caused by a mutation in PIP5K1C, which encodes PIPKI gamma of the phophatidylinsitol pathway. *Am. J. Hum. Genet.* **2007**, *81*, 530–539. [CrossRef]
70. Wang, Y.; Lian, L.; Golden, J.A.; Morrisey, E.E.; Abrams, C.S. PIP5KI gamma is required for cardiovascular and neuronal development. *Proc. Natl. Acad. Sci. USA* **2007**, *104*, 11748–11753. [CrossRef]
71. Porciello, N.; Kunkl, M.; Viola, A.; Tuosto, L. Phosphatidylinositol 4-Phosphate 5-Kinases in the Regulation of T Cell Activation. *Front. Immunol.* **2016**, *7*, 186. [CrossRef] [PubMed]
72. Emerling, B.M.; Hurov, J.B.; Poulogiannis, G.; Tsukazawa, K.S.; Choo-Wing, R.; Wulf, G.M.; Bell, E.L.; Shim, H.S.; Lamia, K.A.; Rameh, L.E.; et al. Depletion of a putatively druggable class of phosphatidylinositol kinases inhibits growth of p53-null tumors. *Cell* **2013**, *155*, 844–857. [CrossRef] [PubMed]
73. Semenas, J.; Hedblom, A.; Miftakhova, R.R.; Sarwar, M.; Larsson, R.; Shcherbina, L.; Johansson, M.E.; Harkonen, P.; Sterner, O.; Persson, J.L. The role of PI3K/AKT-related PIP5K1alpha and the discovery of its selective inhibitor for treatment of advanced prostate cancer. *Proc. Natl. Acad. Sci. USA* **2014**, *111*, E3689–E3698. [CrossRef] [PubMed]
74. Wright, B.D.; Simpson, C.; Stashko, M.; Kireev, D.; Hull-Ryde, E.A.; Zylka, M.J.; Janzen, W.P. Development of a High-Throughput Screening Assay to Identify Inhibitors of the Lipid Kinase PIP5K1C. *J. Biomol. Screen.* **2015**, *20*, 655–662. [CrossRef]

75. Voss, M.D.; Czechtizky, W.; Li, Z.; Rudolph, C.; Petry, S.; Brummerhop, H.; Langer, T.; Schiffer, A.; Schaefer, H.L. Discovery and pharmacological characterization of a novel small molecule inhibitor of phosphatidylinositol-5-phosphate 4-kinase, type II, beta. *Biochem. Biophys. Res. Commun.* **2014**, *449*, 327–331. [CrossRef]
76. Hayakawa, N.; Noguchi, M.; Takeshita, S.; Eviryanti, A.; Seki, Y.; Nishio, H.; Yokoyama, R.; Noguchi, M.; Shuto, M.; Shima, Y.; et al. Structure-activity relationship study, target identification, and pharmacological characterization of a small molecular IL-12/23 inhibitor, APY0201. *Bioorg. Med. Chem.* **2014**, *22*, 3021–3029. [CrossRef]
77. Philippon, H.; Brochier-Armanet, C.; Perriere, G. Evolutionary history of phosphatidylinositol- 3-kinases: Ancestral origin in eukaryotes and complex duplication patterns. *BMC. Evol. Biol.* **2015**, *15*, 226. [CrossRef]
78. Singh, B.; Gupta, R.S. Conserved inserts in the Hsp60 (GroEL) and Hsp70 (DnaK) proteins are essential for cellular growth. *Mol. Genet. Genom.* **2009**, *281*, 361–373. [CrossRef]
79. Gao, B.; Gupta, R.S. Phylogenetic framework and molecular signatures for the main clades of the phylum Actinobacteria. *Microbiol. Mol. Biol. Rev.* **2012**, *76*, 66–112. [CrossRef]

CHAPTER 3

Identification of a conserved 8 aa insert in the PIP5K protein in the *Saccharomycetaceae* family of fungi and the molecular dynamics simulations and structural analysis to investigate its potential functional role

This chapter describes the identification of a conserved 8 amino acid CSI in the core kinase domain of phosphatidylinositol-4-phosphate-5-kinase (PIP5K), which is unique to the *Saccharomycetaceae* family of fungi. PIP5K is a key enzyme in the phosphatidylinositol signaling pathway essential for multiple cellular processes. The results from structural analyses and molecular dynamics (MD) studies provide meaningful insights concerning the mechanism of the interaction of PIP5K protein with membrane lipid bilayers and support the contention that the identified 8 aa conserved insert in *Saccharomyces cerevisiae* plays an important role in facilitating the binding of PIP5K with the membrane surface. My contribution towards the completion of this chapter includes identification of conserved indels, construction of phylogenetic tree, homology modelling studies, molecular dynamics (MD) simulations and analyses of the MD trajectories. I was also involved in writing the manuscript, including preparation of all the figures, tables, and supplementary materials.

Due to limited space, supplementary materials (figures and tables) are not included in the chapter but can be accessed along with the rest of the manuscript at:

Khadka, B and Gupta, R.S. (2017). *Proteins*, 85 (8), 1454-1467.

PROTEINS
STRUCTURE ■ FUNCTION ■ BIOINFORMATICS

Identification of a conserved 8 aa insert in the PIP5K protein in the *Saccharomycetaceae* family of fungi and the molecular dynamics simulations and structural analysis to investigate its potential functional role

Bijendra Khadka and Radhey S. Gupta*

Department of Biochemistry and Biomedical Sciences, McMaster University, Hamilton, Ontario, Canada L8N 3Z5

ABSTRACT

Homologs of the phosphatidylinositol-4-phosphate-5-kinase (PIP5K), which controls a multitude of essential cellular functions, contain a 8 aa insert in a conserved region that is specific for the *Saccharomycetaceae* family of fungi. Using structures of human PIP4K proteins as templates, structural models were generated of the *Saccharomyces cerevisiae* and human PIP5K proteins. In the modeled *S. cerevisiae* PIP5K, the 8 aa insert forms a surface exposed loop, present on the same face of the protein as the activation loop of the kinase domain. Electrostatic potential analysis indicates that the residues from 8 aa conserved loop form a highly positively charged surface patch, which through electrostatic interaction with the anionic portions of phospholipid head groups, is expected to play a role in the membrane interaction of the yeast PIP5K. To unravel this prediction, molecular dynamics (MD) simulations were carried out to examine the binding interaction of PIP5K, either containing or lacking the conserved signature insert, with two different membrane lipid bilayers. The results from MD studies provide insights concerning the mechanistic of interaction of PIP5K with lipid bilayer, and support the contention that the identified 8 aa conserved insert in fungal PIP5K plays an important role in the binding of this protein with membrane surface.

Proteins 2017; 85:1454–1467.

Key words: phosphatidylinositol-4-phosphate-5-kinases; Saccharomyces cerevisiae PIP5K protein; conserved signature insert specific the Saccharomycetaceae family; homology modeling; structural models for human and Saccharomyces cerevisiae PIP5K; molecular dynamics simulations for the PIP5K-membrane interactions.

INTRODUCTION

In eukaryotic organisms, phosphoinositide's (PIs) are important regulators that play an essential role in wide range of cellular processes.[1,2] Different derivatives of phosphoinositides are usually formed by phosphorylation of inositol ring at 3, 4, and 5 positions by various phosphoinositide kinases.[3] One important enzyme in this regard is phosphatidylinositol-4-phosphate-5-kinase (PIP5K), which catalyzes the production of phosphatidylinositol (4,5)-bisphosphate [Ptdlns(4,5)-P2].[4] Among the phosphorylated phosphoinositide's, Ptdlns(4,5)-P2 is one of the key regulators of phosphoinositide signaling pathway.[5] Ptdlns(4,5)-P2 not only acts as a substrate for phosphatidylinositol 3-kinase (PI3K) and receptor-activated phospholipase C (PLC), but it also functions as a second messenger by itself, influencing diverse essential cellular functions.[5–8] In mammalian cells, Ptdlns (4,5)-P2 is predominantly produced by the 'classical PI route' using the enzyme phosphotidylinositol-4-phosphate-5-kinase (PIP5K) to phosphorylate phosphatidylinositol (4) phosphate [Ptdlns(4) P] at the D-5 hydroxyl group of inositol ring. However, Ptdlns (4,5)-P2 is also produced via the enzyme phosphotidylinositol-5-phosphate-4 kinase (PIP4K) through phosphorylation of phosphatidylinositol (5) phosphate [Ptdlns(5) P] at D-4 hydroxyl group of inositol ring.[5]

Additional Supporting Information may be found in the online version of this article.

*Correspondence to: R. S. Gupta; Email: gupta@mcmaster.ca
Received 14 February 2017; Revised 6 April 2017; Accepted 10 April 2017
Published online 13 April 2017 in Wiley Online Library (wileyonlinelibrary.com).
DOI: 10.1002/prot.25306

In mammalian cells, three different isoforms of both PIP4K, as well as PIP5K (viz., PIP5K-α, PIP5K-β, and PIP5K-γ), have been identified.[9,10] The three isoforms of PIP5K share high degree of sequence homology and all exhibit the same enzymatic function, that is, they catalyze the formation of Ptdlns (4,5)-P2 from Ptdlns(4)P.[11,12] The PIP5K homologs also exhibit significant sequence similarities, particularly in the catalytic core regions of the proteins, to PIP4K homologs, but the substrate preference and cellular localization of PIP4K are different.[13,14] The budding yeast *Saccharomyces cerevisiae* contains a single homolog of the PIP5K, which is referred to as the multicopy suppressor of stt4 mutation (MSS4) or PIP5K/MSS4. The PIP5K is a 90 kDa protein, which is primarily localized on the plasma membrane[15,16] and it is required for diverse essential cellular processes, such as membrane trafficking, cell polarity, viability, morphogenesis, cytokinesis and actin cytoskeleton organization.[17,18] Although the N-terminal half of the PIP5K protein shows no sequence similarity to any protein outside of fungi, its C-terminal half is homologous to the mammalian PIP5K and contains the well-conserved catalytic kinase domain.[19] The substrate specificity and plasma membrane targeting of PIP5K is reported to be determined by the activation loop region located within the kinase domain.[14] Despite PIP5K playing an essential role in large numbers of cellular processes, much remains to be elucidated concerning its cellular function. There is no experimentally solved crystal structure for PIP5K from Fungi and the mechanism by which it localizes to the plasma membrane and recognizes or interacts with its substrates is not yet clearly understood. However, structural information for different isoforms of PIP4K from humans is available.[13]

In the past few decades genome sequences have become available from diverse groups of prokaryotic and eukaryotic organisms. Comparative analyses of these genomes are leading to discovery of numerous novel molecular characteristics/markers that are specific for different groups of organisms. Of these molecular markers, conserved signature indels (insertion/deletions) (CSIs), which are specifically present in the protein homologs from a given group of organisms comprise an important category.[20] Due to their evolutionary conservation and specificity for a given group of organisms, the CSIs provide novel tools for evolutionary, genetic and biochemical studies.[21–23] In the present work, analysis of PIP5K protein sequences from various species has identified an 8 aa insert in a conserved region that is uniquely shared by different species belonging to the *Saccharomycetaceae* family of fungi. As no experimentally solved crystal structure of PIP5K from a member of the *Saccharomycetaceae* family was available, we have utilized comparative homology modeling approach to decipher the three-dimensional (3D) structures of PIP5K from *S. cerevisiae* and human (α-isoform). In the modeled structure of *S. cerevisiae* PIP5K, the 8 aa insert forms a positively-charged surface loop, which is present on the same face of the protein as the activation loop of the kinase domain, which based on earlier studies is implicated to play a role in the membrane binding of PIP5K.[24–26] To gain insights into the interactions between PIP5K and membrane, and the role of the 8 aa CSI in the PIP5K, we have carried out molecular dynamics (MD) simulations of *S. cerevisiae* PIP5K structure and different model lipid bilayer membranes. The MD simulation affords an important means for probing the dynamic behavior of interaction of peripheral membrane proteins with lipid bilayer environment.[27–32] The results of the presented structural and biophysical studies have allowed us to model the nature of the PIP5K-membrane association process, and they provide useful insights into the mechanistic and/or structural features of the identified fungi-specific 8 aa conserved insert.

MATERIALS AND METHODS

Sequence alignment and phylogenetic analysis and primary sequence analysis of the target proteins

Blastp searches were carried out on the PIP5K sequence from the human alpha isoform (Genbank ID No. 208431778) and sequences of about 20–25 high scoring homologs from diverse animals and fungi species were retrieved. A multiple sequence alignment of these proteins was created using the Clustal X 1.83 program.[33] Visual examination of this sequence alignment revealed an 8 aa insert in a conserved region that was present in some fungi. Evolutionary significance of this conserved indel was determined by more detailed blastp searches on this indel and its flanking conserved regions, as described in earlier work.[21] A maximum-likelihood phylogenetic tree based on 100 bootstrap replicates of the sequence alignment for PIP5K homologs was constructed using MEGA 6[34] as described in earlier work.[21]

The amino acid sequences of the target protein PIP5K from *S. cerevisiae* (G.I. no. 6320414) and human (G.I. no. 208431778) were retrieved from the Genbank database. The prediction of secondary structure elements of the target protein sequences was carried out using three different servers, PSI-Pred,[35] Jpred3,[36] and PSS-Pred.[37] Genesilico MetaDisorder webserver (http://iimcb.genesilico.pl/metadisorder/) was utilized to predict the intrinsic disorder region in the target protein sequence. MetaDisorder meta-server utilizes the predictions from 13 independent disordered prediction program/servers and calculate the final consensus prediction of disorder.[38]

Homology modeling and electrostatic potential analysis

To search for the appropriate templates for homology modeling, PSI-BLAST[39] searches were carried out on the

protein sequences of *S. cerevisiae* and human PIP5K against the Protein Data Bank (PDB) database.[40] These searches identified three templates corresponding to human PIP4K showing high degree of sequence identity to the target proteins (Supporting Information Table S1). Sequence alignments between the target and the template proteins were created using the *align 2D* module in the Modeller v 9.11[41] and the resulting alignments were carefully analyzed and modified manually. For each PIP5K target protein, 200 models were generated and selected using Discrete Optimized Protein Energy (DOPE) potential score[42] as implemented in Modeller v 9.11. Refinement of the loop region in the modeled structure was carried out using the ModLoop server.[43] The model with the highest DOPE score was then submitted to the ModRefiner program to obtain atomic-level energy minimization and to obtain a model with reliable stereochemistry quality.[44] Homology modeling steps were repeated for generating models based on available single templates as well as multiple templates to improve the overall quality of the final model. The quality check of the final structure model was carried out using three different independent servers; RAMPAGE,[45] ProsA,[46,47] and QMEAN[48] server. The modeled structures were visualized and analyzed using the molecular visualization program PyMol (http://www.pymol.org).

The electrostatic potential surface (EPS) of modeled PIP5K proteins was calculated using the PBEQ-Solver server (http://www.charmm-gui.org/?doc=input/pbeqsolver). PBEQ-Solver is a web-based graphical user interference that allows users to solve the Poisson–Boltzmann (PB) equation and interactively visualize the electrostatic potential of biomolecules.[49] The electrostatic potential scale was set to from −10 Kcal/mol to +10 Kcal/mol and visualized using PyMol.

Molecular dynamic (MD) simulation of the interaction of PIP5K with model lipid membrane bilayers

Molecular dynamics simulations were performed using GROMACS 5.1.2 software[50,51] with a GROMOS96 53A7 force field, and the SPC water model.[52] The system was built using the previously fully hydrated and equilibrated 3D coordinates of POPC (palmitoyl-oleoyl phosphatidylcholine) and DPPC (dipalmitoyl-phosphatidylcholine) bilayers[53,54] and the lipid parameters were taken from Berger et al.[55] The membranes composed of both POPC and DPPC lipids represent the eukaryotic membrane environment and of these POPC lipids are highly abundant in yeast cell membrane.[56] Particle Mesh Ewald method was utilized to calculate the electrostatic interactions[57,58] and the short-range repulsive and attractive dispersion interactions were described using Lennard–Jones potential that was cutoff at 1.2 nm. The bond lengths and the angles in the water molecules were constrained using SETTLE algorithm,[59] and LINICS[60] was used to constrain all other bonds. After 50,000 steps of energy minimization using steepest descents method, the system was gradually heated from 0 to 323 K in the canonical (NVT) ensembles >100 ps of MD using V-rescale thermostat method.[61] The system was further equilibrated for 1 ns in the isothermal–isobaric (NPT) ensemble using the Nose–Hoover thermostat and Parrinello–Rahman barostat.[62,63] The final production run was continued for 100 ns. In our simulation study, two different model membranes were utilized to analyze the role of the 8 aa conserved insert in membrane interaction. In the first system, the structural model of the *S. cerevisae* PIP5Ks, either containing-the 8 aa insert and lacking this insert, were initially positioned at a distance >10 Å away from the surface of POPC model lipid membrane bilayers, with positively-charged flat surfaces facing toward the membrane surface. The structure model of *S. cerevisiae* PIP5K protein lacking the CSI was generated using the homology modeling approach as described above. In the second system, similar studies were carried out with the PIP5Ks containing the insert or lacking the insert using DPPC lipid bilayer. The initial relative orientations of the PIP5K-containing CSI and PIP5K–lacking CSI with respect to the membrane surface were same for POPC and DPPC membrane lipid bilayers. Appropriate counterions were added by replacing water molecules to neutralize the system. In total, there were 244853 atoms in the simulation box for the system with CSI-containing PIP5K and 245133 atoms for the system with CSI-lacking PIP5K with POPC membrane, and 242059 atoms in the simulation box for the system with CSI-containing PIP5K and 242333 atoms for the system with CSI-lacking PIP5K with DPPC membrane. The secondary structure assessment was carried out using DSSP program.[64] The analyses of the simulation trajectories were performed using tools available in the GROMACs suite.[50,61] All of the MD simulation runs were carried out using our local GROMACS certified graphical processing unit (GPU)-accelerated high-performance computing system obtained from EXXACT Corporation.[65]

RESULTS

Importance of evolutionary conserved indels in protein sequences and identification of a conserved insert in PIP5K homologs specific for the fungi (Saccharomycetaceae)

Inserts or deletions in conserved regions of proteins, which are restricted to homologs from particular groups of organisms, provide useful molecular tools for genetic, biochemical, and evolutionary studies.[21–23,66] Earlier works on conserved signature indels (CSIs) in multiple proteins have established that that such indels play important functional, and often essential, roles in the organisms where they are found.[22,67,68] In the present

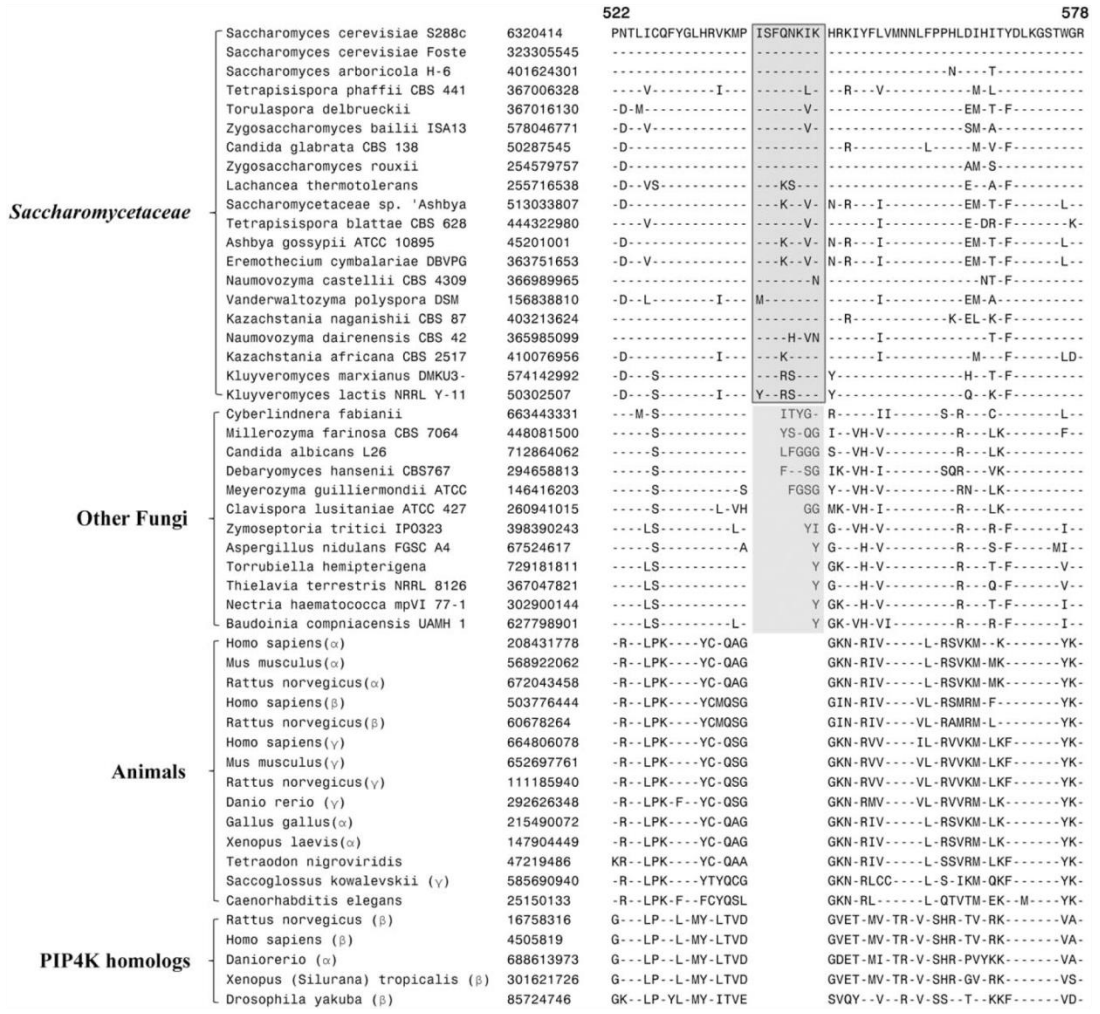


Figure 1

Partial sequence alignment of phosphatidylinositol-4-phosphate 5-kinase (PIP5K) homologs showing a 8 aa conserved insert that is uniquely shared by various species from the family *Saccharomycetaceae*, but not found in any animal species. Smaller inserts that might be specific for other fungi are also present in this position. This insert is also not present in any PIP4K homologs. Sequence information is shown here for only a limited number of species/homologs, however, the characteristics of the conserved insert noted above apply to all. The dashes (–) in the alignment denote identity with the amino acid shown on the top line. The Genbank ID numbers of different sequences are shown in the second column.

work, we describe the characteristics of a 8 aa conserved insert in the PIP5K homologs that is specific for a particular group of fungi (Fig. 1). Detailed blast searches on the sequence region containing this insert showed that within fungi, all of the species belonging to the family *Saccharomycetaceae*, which includes *S. cerevisiae*, contained an 8 aa insert, whereas smaller inserts were also present in some other fungi. Further as seen from Figure 1, the PIP5K homologs from different animal species (sequence information is shown for some species) do not contain an insert in this position. Figure 1 also includes sequence information for representative sequences of some PIP4K homologs and this insert is also not found in the PIP4K homologs.

The evolutionary significance of this conserved insert was further investigated by constructing a phylogenetic tree for the PIP5K homologs from various fungi and a limited number of animal species, as well as some PIP4K homologs (Fig. 2). In this tree, which was rooted using the sequences for the PIP4K proteins, PIP5K homologs

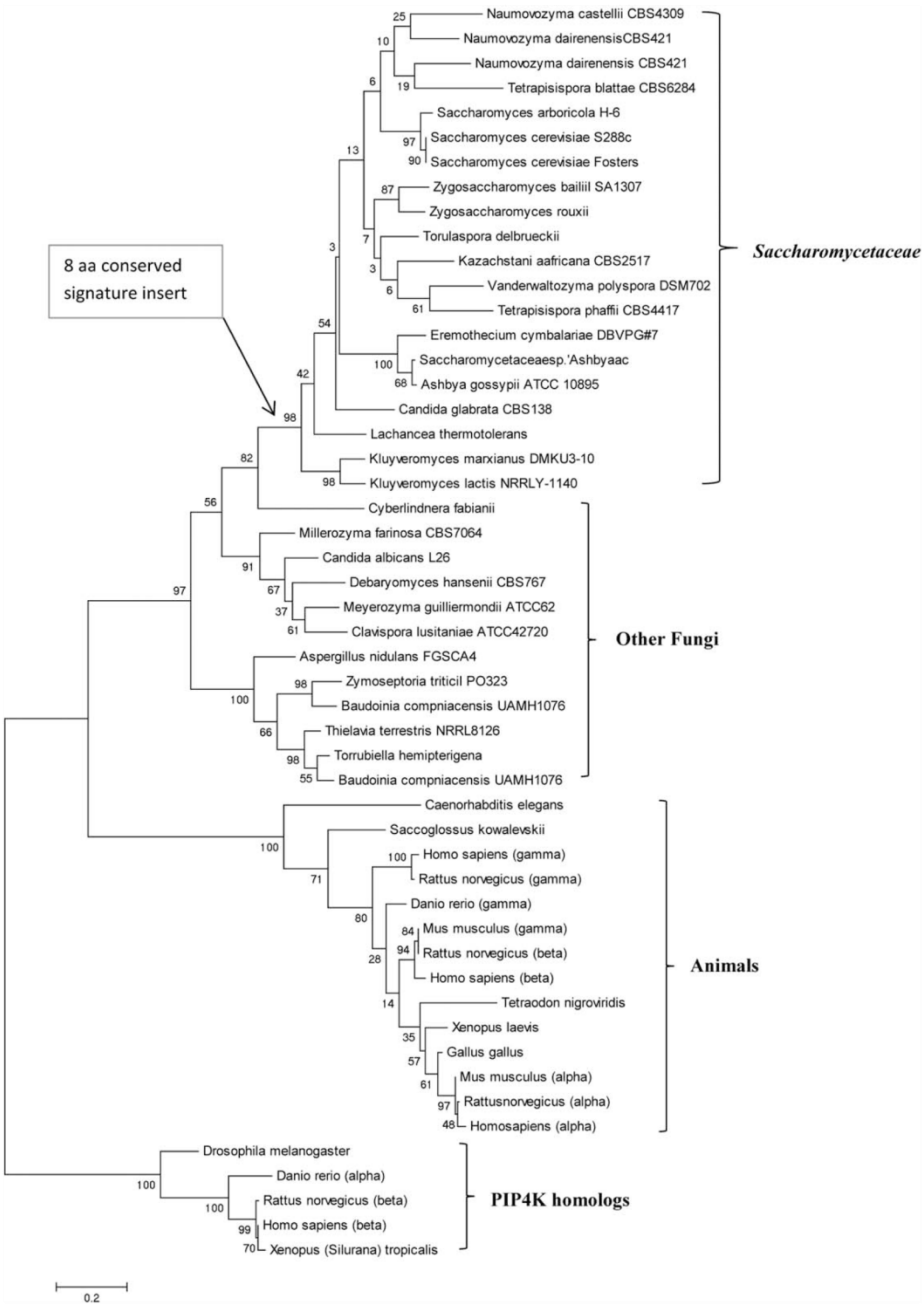


Figure 2

A bootstrapped maximum-likelihood tree based upon PIP5K protein sequences. This tree was rooted using the sequences for the PIP4K homologs. All of the species containing the 8 aa insert (indicated by an arrow) form a strongly supported clade, which corresponds to the *Saccharomycetaceae* species.

from mammalian species branched distinctly from those of the fungi. Three different isoforms of PIP5K, which are found in mammals (viz., rat, mouse and guinea pig) and other cold-blooded animals, also formed separate clusters, as also observed in earlier work.[69] The PIP5K homologs from fungal species, which are part of our analysis, also formed a number of distinct clades. Interestingly, all of the species containing the 8 aa insert formed a strongly supported clade in the tree that corresponded to the *Saccharomycetaceae* family of fungi. Most of the species which contained 4–5 aa inserts in this position are part of the *Debaryomycataceae* family of fungi. Since fungal kingdom is very large comprising of >70 different families,[20,70,71] it is possible that inserts of different lengths in this position could be characteristics of specific families/orders of fungi. Although this aspect needs to be further studied, results presented here strongly indicate that the 8 aa insert in the PIP5K protein is a distinctive characteristic of the *Saccharomycetaceae* family of fungi. The members of this family contain large numbers of important fungi, including the widely studied model organism *S. cerevisiae* that are of considerable environmental and economic importance.[72] Hence, our further work has focused on understanding the structural and functional significance of this 8 aa conserved insert.

Sequence analysis of phosphotidylinositol-4-phosphate-5-kinase (PIP5K)

Analysis of the PIP5K protein from *S. cerevisiae* using the Genesilico MetaDisorder web server indicated that both the 8 aa conserved insert in this protein, as well as the kinase activation loop in this sequence, which is a shared characteristic of the PIP5K family of proteins, are present in the ordered regions of the protein (Supporting Information Fig. S1). We have also carried out secondary structure element analysis of the PIP5K protein sequence using three different servers; PSI-Pred,[35] JPred,[36] and PSSPred.[37] The results of these analyses indicated that the amino acids corresponding to the 8 aa conserved insert are predicted to form a coil/loop region and its flanking residues are predicted to form beta strands [Fig. 3(A)]. Similar secondary structure elements analysis of the amino acids residues from the activation loop indicates the presence/formation of a helix region within the activation loop of the PIP5K protein [Fig. 3(A)]. The results from secondary structure analysis are in broad agreement with biochemical and nuclear magnetic NMR studies on activation loop of PIP5K.[25]

Homology modeling of *S. cerevisiae* and human PIP5K proteins

The absence of any experimentally solved structure for the *Saccharomycetaceae* PIP5K severely restrains our ability to understand the functional significance of the conserved insert present in this protein. Hence, we have used the homology modeling technique to generate models of the PIP5K proteins from *S. cerevisiae* and human to gain insights concerning the structural characteristics of the 8 aa insert present in the *Saccharomycetaceae* homologs. Suitable templates for modeling of these proteins were obtained by performing PSI-BLAST searches with the *S. cerevisiae* and human homologs against the PDB database. As illustrated in Supporting Information Table S1, these searches identified three structures showing significant sequence identity with higher structure resolution (<3.0 Å resolution) that could act as potential templates for the target proteins. All three of these crystal structures corresponded to the different isoforms (alpha, beta, and gamma) of the PIP4K proteins from *Homo sapiens*. Due to low structural resolution and missing structural information, PIP5K1A protein from zebrafish[24] was not included as a candidate for template. Except for these four proteins, no other template showing significant similarity to the target proteins was identified. Sequence similarity of the *S. cerevisiae* PIP5K protein to the template sequences was restricted to the C-terminal half of the protein (that is, residues 393–753). The N-terminal half of the *S. cerevisiae* PIP5K protein (residue ranges from 1 to 392), while conserved in fungi, showed no significant sequence or structural homology to any other protein in the protein data bank (PDB) or the NCBI database. This region was also predicted to be highly disordered and hence it was excluded from modeling studies.

Homology models for the C-terminus half of the *S. cerevisiae* PIP5K protein sequence (393–753 aa) were generated using the three available high-resolution templates as described in the "Methods" section. Homology modeling using multiple templates have been shown to improve the quality of the generated models.[73,74] The multiple sequence alignment between the target protein and the templates used for homology modeling is provided in Supporting Information Figure S2. Structural refinement of the best model and its validation was carried out using ModRefiner energy minimization server and the RAMPAGE server, respectively as described in the Methods section. The Ramachandran plot obtained from RAMPAGE shows that 98% of the residues are in allowed region, 1.7% residues in the favored region and 0.3% residues (a single residue GLN44) in the disallowed region. Residues constituting the 8 aa conserved insert and its flanking region were all found to be in allowed region [Supporting Information Fig. S3(A)]. The Quality Model Energy Analysis (QMEAN) score of >0.5 indicates that the model generated is of good quality. Furthermore, *Z* score predicted using ProSA-web server provides an indication of the overall quality of a model and it is commonly used to ensure the compatibility of the *Z* scores range between input target protein structures and native protein of similar size.[46,47] The *Z* score for

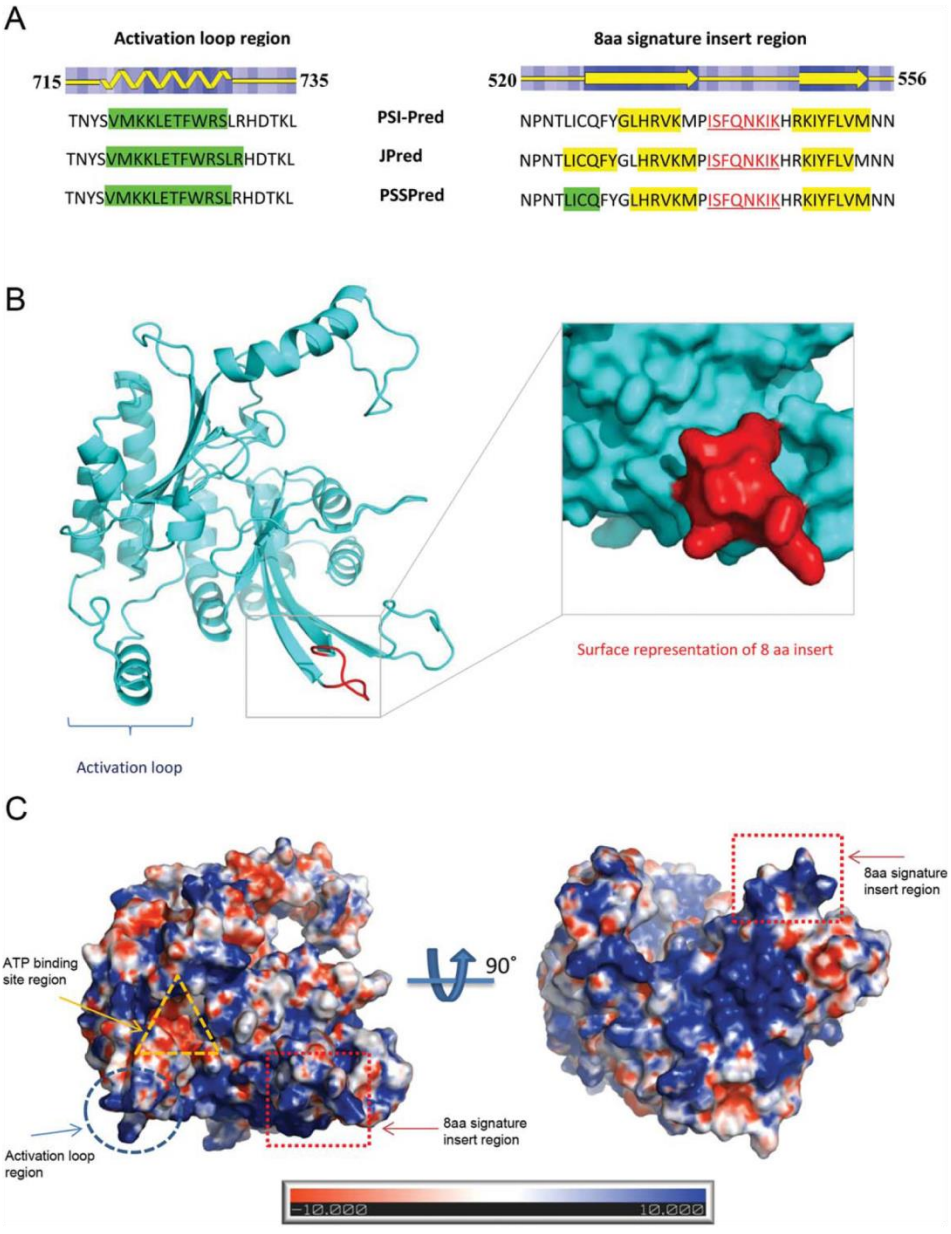


Figure 3

(**A**) Secondary structure predictions for the PIP5K sequence from *S. cerevisiae* for the activation loop (residues 715–735) region and the 8 aa CSI-containing region (residues 520–556). The insert residues are underlined and colored red. Residues predicted to form helix and strand are highlighted by green and yellow colors. (**B**) Cartoon representation of homology model of *S. cerevisiae* PIP5K. The inset on the right shows surface representation of part of the model structure (cyan) with the 8aa insert region exposed on the surface in red color. (**C**) Surface representation of PIP5K from *S. cerevisiae* showing the distribution of charge surface. The location of the region formed by the 8 aa CSI is indicated by red dash boxes, the region formed by activation loop is indicated by blue dash circles and the region involved in ATP binding indicated is shown by yellow triangle. The electronegative potential scale was set to −10 Kcal/mol to +10 Kcal/mol. The positively charged surface is shown in blue whereas negatively charged surface is shown in red. The view on the right shows PIP5K charge distribution after 90° rotation revealing the highly dense positively charged flat surface. Electrostatic Potential surface (EPS) distribution was calculated using PBEQ-Solver server (http://www.charmm-gui.org/?doc = input/pbeqsolver).

PIP5K model was −7.61 [Supporting Information Fig. S3(B)] which is comparable to the *Z* score predicted for the template structures, used as a reference [Supporting Information Fig. S4(B) and Table S2]. The ProSA plots were also generated for both the modeled protein and the templates to display the local quality of model/template by plotting the knowledge-based energies as a function of amino acid sequence position [Supporting Information Fig. S4(AB)]. Overall, the results predicted using different servers indicates that the PIP5K model structure was of good quality and thus can be reliably used for further analysis. The model for the human PIP5K protein (residue ranges from 69 to 432 aa) was generated and validated in a similar manner and its validation characteristics are similar to that for the PIP5K model [Supporting Information Table S2 and Supporting Information Fig. S4(A)], indicating that this model was also of high quality.

Structural comparisons of the *S. cerevisiae* PIP5K model with that of the human PIP5K protein model and the solved structures of the template proteins (i.e., different isoforms of human PIP4K) are shown in Supporting Information Figure S5. There are two main differences seen in the structure of the modeled proteins with those of the solved template structures. First, unlike the solved structures of PIP4K proteins, where the structure of the kinase domain activation loop was not determined, this region is now shown to contain an alpha helix, as predicted by secondary structure analysis. Recently, Liu et al.[25] have solved the structure of the activation loop of PIP5K using nuclear magnetic NMR studies and their results also show the formation of an amphipathic helix within the activation loop upon the interaction with the membrane.[25] Second, the structure of the PIP5K protein from *S. cerevisiae* differed from its mammalian counterpart and from the structure of the PIP4K protein in the region where the 8 aa conserved insert is found in the yeast protein. The structural superimposition of the *S. cerevisiae* PIP5K homology model with the recently solved crystal structure of the catalytic domain of zebrafish PIP5K1A at 3.31 Å resolution[24] shows significant structural similarity with RMSD value of 1.48 Å. The amino acid residues corresponding to the 8 aa insert found in the *Saccharomycetaceae* homologs formed an extended loop exposed to the protein surface, which is absent in the structures of the mammalian PIP5K or PIP4K homologs [Fig. 3(B) and Supporting Information Fig. S5]. Although the loop corresponding to the 8 aa conserved insert is located in the conserved catalytic domain of the PIP5K,[4,5] it is not in the immediate proximity of the activation loop or the phosphate/ATP-binding site.

Electrostatic potential surface of PIP5K

The role of electrostatic interaction between the positively charged surface of proteins and the negative part of the headgroup of lipid bilayers in cellular localization has now been clearly elucidated in several studies.[26,75–77] The crystal structures of PIP4Ks, which were used as templates for homology modeling of the target PIP5K in our study, and the recently solved crystal structure of zebrafish PIP5K1A all contain a positively charged flat surface region that is indicated to interact with negative part of the headgroup of lipid membrane bilayer, using electrostatic interaction.[13,75] Although the distribution of positive charges differ between PIP5K1A and PIP4K, the orientation of the catalytic sites relative to the positively charged flat membrane-binding surface, which is a fundamental feature for interfacial catalysis, is conserved.[24] In view of this, to investigate the putative membrane binding sites on the surface of the modeled PIP5K protein, the electrostatic potentials surface was calculated. As shown in Figure 3(C), the flat surface formed by beta sheets in *S. cerevisiae* PIP5K contains highly dense positively charged region. The presence of the two positively charged Lys-residues in the 8 aa conserved insert further contributes to the positively charged surface patch in the *S. cerevisiae* PIP5K. Previously, Fairn et al.[26] have shown the presence of similar positive surface electrostatic potential on the equivalent surface of structural model of three different isoform of mammalian PIP5K (PIP5Kα, PIP5Kβ, and PIP5Kγ), and recently Hu et al.[24] have reported similar features in the crystal structure of PIP5K1A. These studies have suggested a role for the positive charged patch on the surface of PIP5K in directing the protein to the anionic head groups in lipid membranes.[24,26] Essentially, the *S. cerevisiae* PIP5K models have electrostatic potential profiles similar to that of mammalian PIP4Ks and PIP5K. The 8 aa conserved insert in *S. cerevisiae*, due to its presence on the same face of the protein as the positively charged surface patch, and its contribution to the overall surface charge of the protein is thus likely to play a role in facilitating the interaction of the PIP5K protein with the membrane.

Molecular dynamic simulation of PIP5K-membrane system: analysis of the PIP5K conformational stability and flexibility

We have used the MD simulation approach to probe the interaction of PIP5K with model membranes to gain insights into the functional role of the 8 aa fungi-specific insert. In the kinase proteins, such as PIP5K, several features which interact with the surface of membrane bilayer are known to be well conserved.[24,26,78] The interaction of these proteins with membrane surface is primarily driven by the electrostatic interaction through their flattened positively charged surface and the anionic portions of phospholipid molecules in the membranes.[24,26,78] However, the extent of dynamic mechanism of interaction between the *Saccharomycetaceae* PIP5K and membrane bilayer has not been fully

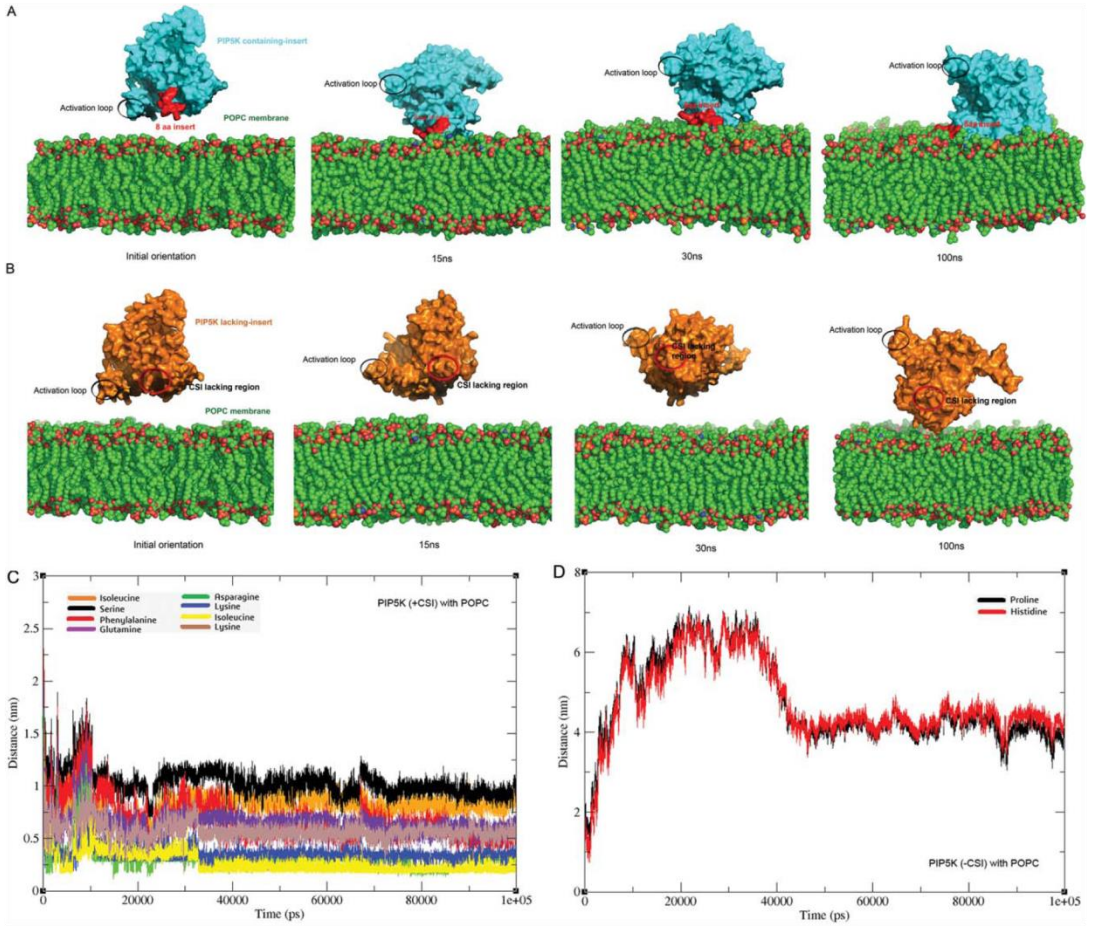


Figure 4

(**A**) Snapshots from simulation studies showing the binding interaction of *S. cerevisiae* PIP5K to the POPC lipid bilayer membrane. The initial orientation and the snapshots from 15, 30, and 100 ns simulation periods are shown. The CSI-containing PIP5K is shown in surface representation (green) and the 8 aa CSI is shown as red surface. The position of the activation loop is marked with a black circle. The POPC lipid molecules are shown as ball and stick models (red and green color). (**B**) The binding interaction of the CSI-lacking PIP5K to POPC lipid bilayer at various time during the simulation run. The CSI-lacking PIP5K is shown as orange surface and the region where the 8 aa CSI is found is shown by red circle. (**C**) The time evolution of the minimum distance between the residues from the 8 aa CSI and POPC membrane bilayer during 100 ns simulation for the CSI-containing PIP5K. Specific residues from the 8 aa CSI are shown as ILE 147 (Orange), SER148 (Black), PHE149 (Red), GLU150 (Purple), ASP151 (Green), LYS152 (Blue), IIL153 (Yellow), and LYS154 (Grey). (**D**) Minimum distance between the residues from the position where the 8 aa CSI is found (PRO146 (black) and HIS147 (red) and POPC membrane bilayer during 100 ns simulation for the CSI-lacking PIP5K.

understood. To study the association of the *S. cerevisiae* PIP5K with model membranes and the role of the 8 aa conserved in this interaction, MD simulations of the PIP5K was performed in two different model lipid environments consisting of POPC and DPPC lipid bilayers. The details of the simulation set up used are described in Method section.

To ensure the global stability of the PIP5K and the 8 aa insert during simulation in different membrane environments the root mean square deviation (RMSD) analysis was initially carried out. The RMSD values were computed for Cα of both insert-containing and insert-lacking PIP5K models to predict their overall stability over the course of the simulation. During the simulation with POPC and DPPC, the RMSD values for both CSI-containing and –lacking PIP5K did not fluctuate significantly indicating that the PIP5Ks remains stable throughout the simulation [Supporting Information Figs.

S6(A,B) and S7(A,B)]. Likewise, the RMS fluctuation (RMSF) of each α-carbon backbone atoms of the amino acid residues in the PIP5K averaged over the simulation time in POPC and DPPC membrane environment was also calculated [Supporting Information Figs. S6(C,D) and S7(C,D)]. The high peaks in these figures indicate that these regions in PIP5K are flexible. The position of residues indicated in the plots corresponds to the C-terminal half of the PIP5K with the residue position 393–753. Results of these analysis indicate that the residues from the protein forming the positively-charged flat surface showed the least amount of movement except for the loop regions, where higher peaks were observed. The other high peaks were observed near the C-terminal region that corresponds to highly divergent and disordered sequence. The region from the PIP5K that contains the 8 aa conserved insert showed RMSF values that are typical of exposed surface loop region. The stability of the secondary structure elements of PIP5K during the simulation period was also calculated. As shown in Supporting Information Figure S8(A,B), the secondary structure elements of the 8 aa insert forming surface loop appeared very stable throughout the simulation. Likwise, the alpha-helical structure present within the activation loop also maintained stability during the simulation.

Analysis of the PIP5K-lipid bilayer interactions

Visual examination of the interaction of the PIP5K protein structures with the membrane bilayers in our simulations revealed that the 8 aa insert interacted with the membrane surfaces throughout the 100 ns period, and the binding of the proteins with the lipid bilayers progressively increased as a function of time. The major events in our simulation of the interaction of PIP5K with the POPC and DPPC membrane surfaces are illustrated in the Figure 4 and Supporting Information Figure S9. Figure 4(A,B) show series of snapshots showing the initial orientation of the PIP5K protein containing or lacking the CSI relative to the membrane surface, and their orientations at 15, 30, and 100 ns simulation periods. During the simulation, the insert-containing PIP5K initially interacted with the POPC and DPPC membrane models via a surface exposed loop formed by the 8 aa conserved insert (residue position 147–154) and subsequently the positively charged flat surface of the PIP5K associated with membrane surface to form a more stable association. This was confirmed by measuring the distance between residues from the 8 aa insert and membrane bilayer surface. In the POPC membrane system, for the CSI-containing PIP5K, distance of the protein from the membrane bilayer decreased from initial greater than10 to <5 Å over within 2 ns and then it showed no significant change over the course of 100 ns simulation [Fig. 4(C)]. In contrast, for the protein lacking the 8 aa CSI, no interaction with the POPC bilayer was observed at 15 and 30 ns, but at 100 ns some binding of the protein to the membrane was observed [Fig. 4(B,D)]. However, this binding was in a different orientation than that observed for the CSI-containing protein and the region where the CSI is found remained distal from the membrane over the simulation period.

In the DPPC lipid system, although both the CSI-containing and CSI-lacking proteins were found to bind to the membrane, their binding configurations were different [Supporting information Fig. S9(A,B)]. In this system, the distance of the CSI-containing region of the protein from the membrane bilayer decreased to <5 Å within 2 ns and showed no further change. However, for the protein lacking the CSI, the distance from PRO146 and HIS 147, which correspond to the residues that flank the CSI, and the membrane bilayer showed considerable fluctuation over the course of 100 ns simulation [Supporting Information Fig. S9(C,D)]. For the protein lacking the insert, activation loop region was also found to interact with the DPPC membrane bilayer [Supporting Information Fig. S9(B)].

A more detailed analysis of the interaction between protein and membrane shows that the residues from the 8 aa insert viz., ASP151 (corresponds to residue 543), LYS152 (corresponds to residue 544), ILE153 (corresponds to residue 545), and LYS154 (corresponds to residue 546) formed a direct interaction with the anionic head groups of POPC and DPPC membrane lipid bilayer by forming hydrogen bonds. Less extensive nonbonded contacts were observed between the residues ILE147 (corresponds to residue 539), PHE149 (corresponds to residue 541), GLN150 (corresponds to residue 542) and the membrane surface.

DISCUSSION

This work reports an 8 aa conserved insert in the enzyme PIP5K which is specific for the *Saccharomycetaceae* family of fungi. The enzyme PIP5K due to its central role in the generation of phosphoinositide, Ptdlns(4,5)-P2, a key molecule in the phosphoinositide signalling pathway, plays an essential role in controlling diverse cellular processes. Inserts of other lengths are also present in the same position in other fungi, which might be specific for other groups of fungi. Earlier work on CSIs provides evidence that the genetic changes represented by them are essential for the groups of organisms where they are found.[22,68,79,80] Removal of these CSIs or any significant changes in them are incompatible with cellular growth indicating that these CSIs play important functional roles.[22] Thus, the 8 aa CSIs in PIP5K, which is specific for the *Saccharomycetaceae* homologs is also expected to play an important function in these fungi.

Most studied CSIs in protein structures are present in the surface loops of proteins, which are predicted to play important roles in mediating novel protein–protein or protein–ligand interactions that are specific for the CSI-containing organisms.[67,68,80–84] Our comparison of the modeled structures of PIP5K proteins from *S. cerevisiae* and humans reveal that the 8 aa insert in the *S. cerevisiae* PIP5K is also present in an exposed surface loop. This loop is located away from the activation loop as well as ATP-binding site and thus it is unlikely that this loop, or the residues that are part of the 8 aa insert, play a direct role in the catalytic activity of the protein, which is a conserved characteristic of the PIP5K family of proteins. However, this loop is present on the same face of the protein as the activation loop and due to the presence of two lysine residues in the 8 aa insert, this loop has a positively charged character. The essential role of positively charged surface and highly conserved residues from protein kinases in membrane binding has been demonstrated in several previous modeling and mutagenesis studies.[14,85–87]

MD simulation technique was used in this work to investigate the interaction of the CSI-containing and lacking PIP5K proteins with model membrane bilayers. The results from these studies support the view that the residues from the 8 aa conserved insert, through nonspecific electrostatic interaction, help the *S. cerevisiae* PIP5K protein to anchor and associate with the model membrane lipid bilayer surfaces. Our simulations predicts a binding orientation for the CSI-containing PIP5K protein which is similar to that proposed for the mammalian PIP5K.[24] In contrast, the binding configuration of the CSI-lacking PIP5K to the membrane was quite different indicating that the presence of the CSI affects the binding of the protein to the membrane. The residues from the CSI likely affect the overall binding interaction of the protein by playing a role in the anchoring of PIP5K to membrane surface by binding to negative part of the zwitterionic headgroup of lipid membrane bilayers. Additionally, the presence of this insert could also facilitate closer apposition and proper orientation of the *Saccharomycetaceae* PIP5K protein with membrane surface, and this could lead to a more productive conformation of the protein. The surface loops in several other peripheral proteins have also been shown to play an important role in interaction with membrane surface.[30,87–89]

It should be noted that based upon our study it is difficult to infer the absolute final stable orientation of *S. cerevisiae* PIP5K on the membrane bilayer. The *S. cerevisiae* PIP5K protein studied in the present work is lacking the N-terminal half of the protein, which shows no sequence similarity to any known protein. Thus, it is difficult to predict what role the missing N-terminal region of the protein may be playing in the membrane interaction of the fungal PIP5K protein. It should also be noted that the mammalian PIP5K protein functions as a dimer[12] and the dimeric form may be necessary for maintaining the protein in its proper conformation. However, there is no information available concerning the functional form of the fungal PIP5K protein and whether it also functions as a dimer. Additionally, the MD simulation in the present work was carried out in a very simplified system and for relative short period of time. Simulations of protein–bilayer interactions often require time scale of several hundred nanoseconds for full sampling.[90] Given the complexity of fungal membrane structure,[56,91] future multiscale MD simulation studies, with a more complex mixture of lipids-particularly ergosterol, will be useful to provide additional information regarding the proper binding orientation of the *S. cerevisiae* PIP5K to the membrane.

Despite the above limitations, the results presented strongly suggest that the surface loop formed by the 8 aa conserved insert in the *Saccharomycetaceae* PIP5K homologs plays an important role in the binding or anchoring of these proteins to the fungal membrane surface. Based on the specificity of the 8 aa insert for the *Saccharomycetaceae* PIP5K, it is likely that the observed binding interaction is specific for this family of fungi and they are not found in other eukaryotic organisms. It is of interest to note that the fungal cell walls contain polysaccharides and instead of cholesterol, ergosterol is present in the fungal membranes.[91,92] These characteristics will affect the binding of PIP5K with the membrane and they can also provide a plausible explanation for the presence of inserts of different lengths in this position in other fungi. However, the possibility that the role played by this conserved insert is played by some other protein in higher organisms cannot be excluded. Further genetic and biochemical studies in model organisms such as *S. cerevisiae*[15,26,93] should prove very helpful in understanding the functional significance of these conserved inserts. Lastly, it should be noted that the proteins involved in phosphoinositide signalling pathway, due to their central role in controlling diverse cellular processes, provide potential drug targets.[69,94,95] In this context, the observed specificity of the 8 aa insert for the *Saccharomycetaceae* family of fungi, and the predicted essential function of this insert, the identified CSI in the PIP5K protein could serve as a potential target for the development of novel agents targeting this group of fungi.

ACKNOWLEDGMENTS

This work was supported by a research grant (No. 249924) from the Natural Science and Engineering Research Council of Canada. We thank Dr. Richard Epand for helpful discussions regarding PIP Kinases.

REFERENCES

1. Kutateladze TG (2010) Translation of the phosphoinositide code by PI effectors. Nat Chem Biol 6:507–513.

2. Majerus PW (1992) Inositol phosphate biochemistry. Annu Rev Biochem 61:225–250.
3. Wuttke A, Sagetorp J, Tengholm A (2010) Distinct plasma-membrane PtdIns(4)P and PtdIns(4,5)P2 dynamics in secretagogue-stimulated beta-cells. J Cell Sci 123:1492–1502.
4. Loijens JC, Boronenkov IV, Parker GJ, Anderson RA (1996) The phosphatidylinositol 4-phosphate 5-kinase family. Adv Enzyme Regul 36:115–140.
5. van dBI, Divecha N (2009) PIP5K-driven PtdIns(4,5)P2 synthesis: regulation and cellular functions. J Cell Sci 122:3837–3850.
6. Logan MR, Mandato CA (2006) Regulation of the actin cytoskeleton by PIP2 in cytokinesis. Biol Cell 98:377–388.
7. Mao YS, Yin HL (2007) Regulation of the actin cytoskeleton by phosphatidylinositol 4-phosphate 5 kinases. Pflugers Arch 455:5–18.
8. Yin HL, Janmey PA (2003) Phosphoinositide regulation of the actin cytoskeleton. Annu Rev Physiol 65:761–789.
9. Ishihara H, Shibasaki Y, Kizuki N, Katagiri H, Yazaki Y, Asano T, Oka Y (1996) Cloning of cDNAs encoding two isoforms of 68-kDa type I phosphatidylinositol-4-phosphate 5-kinase. J Biol Chem 271: 23611–23614.
10. Ishihara H, Shibasaki Y, Kizuki N, Wada T, Yazaki Y, Asano T, Oka Y (1998) Type I phosphatidylinositol-4-phosphate 5-kinases. Cloning of the third isoform and deletion/substitution analysis of members of this novel lipid kinase family. J Biol Chem 273:8741–8748.
11. Epand RM (2012) Recognition of polyunsaturated acyl chains by enzymes acting on membrane lipids. Biochim Biophys Acta 1818: 957–962.
12. Shulga YV, Anderson RA, Topham MK, Epand RM (2012) Phosphatidylinositol-4-phosphate 5-kinase isoforms exhibit acyl chain selectivity for both substrate and lipid activator. J Biol Chem 287:35953–35963.
13. Rao VD, Misra S, Boronenkov IV, Anderson RA, Hurley JH (1998) Structure of type IIbeta phosphatidylinositol phosphate kinase: a protein kinase fold flattened for interfacial phosphorylation. Cell 94: 829–839.
14. Kunz J, Fuelling A, Kolbe L, Anderson RA (2002) Stereo-specific substrate recognition by phosphatidylinositol phosphate kinases is swapped by changing a single amino acid residue. J Biol Chem 277: 5611–5619.
15. Guillas I, Vernay A, Vitagliano JJ, Arkowitz RA (2013) Phosphatidylinositol 4,5-bisphosphate is required for invasive growth in *Saccharomyces cerevisiae*. J Cell Sci 126:3602–3614.
16. Gary JD, Wurmser AE, Bonangelino CJ, Weisman LS, Emr SD (1998) Fab1p is essential for PtdIns(3)P 5-kinase activity and the maintenance of vacuolar size and membrane homeostasis. J Cell Biol 143:65–79.
17. Desrivieres S, Cooke FT, Parker PJ, Hall MN (1998) MSS4, a phosphatidylinositol-4-phosphate 5-kinase required for organization of the actin cytoskeleton in *Saccharomyces cerevisiae*. J Biol Chem 273:15787–15793.
18. Homma K, Terui S, Minemura M, Qadota H, Anraku Y, Kanaho Y, Ohya Y (1998) Phosphatidylinositol-4-phosphate 5-kinase localized on the plasma membrane is essential for yeast cell morphogenesis. J Biol Chem 273:15779–15786.
19. Anderson C (2015) How to give a killer presentation, Vol. 91. Harvard: Harvard Business Review, pp 121–125.
20. Gupta RS (1998) Protein phylogenies and signature sequences: a reappraisal of evolutionary relationships among archaebacteria, eubacteria, and eukaryotes. Microbiol Mol Biol Rev 62:1435–1491.
21. Gupta RS (2014) Identification of conserved indels that are useful for classification and evolutionary studies. In: Goodfellow M, Sutcliffe IC, Chun J, editors. Bacterial taxonomy, methods in microbiology, Vol. 41. London: Elsevier, pp 153–182.
22. Singh B, Gupta RS (2009) Conserved inserts in the Hsp60 (GroEL) and Hsp70 (DnaK) proteins are essential for cellular growth. Mol Genet Genomics 2009;281:361–373.
23. Bhandari V, Naushad HS, Gupta RS (2012) Protein based molecular markers provide reliable means to understand prokaryotic phylogeny and support Darwinian mode of evolution. Front Cell Infect Microbiol 2012;2:98.
24. Hu J, Yuan Q, Kang X, Qin Y, Li L, Ha Y, Wu D (2015) Resolution of structure of PIP5K1A reveals molecular mechanism for its regulation by dimerization and dishevelled. Nat Commun 2015;6:8205.
25. Liu A, Sui D, Wu D, Hu J (2016) The activation loop of PIP5K function as a membrane sensor essential for lipid substrate processing. Sci Adv 2:e1600925.
26. Fairn GD, Ogata K, Botelho RJ, Stahl PD, Anderson RA, De Camilli P, Meyer T, Wodak S, Grinstein S (2009) An electrostatic switch displaces phosphatidylinositol phosphate kinases from the membrane during phagocytosis. J Cell Biol 187:701–714.
27. Biggin PC, Bond PJ (2008) Molecular dynamics simulations of membrane proteins. Methods Mol Biol 443:147–160.
28. Cui H, Ayton GS, Voth GA (2009) Membrane binding by the endophilin N-BAR domain. Biophys J 97:2746–2753.
29. Psachoulia E, Sansom MS (2009) PX- and FYVE-mediated interactions with membranes: Simulation studies. Biochemistry 48:5090–5095.
30. Lumb CN, He J, Xue Y, Stansfeld PJ, Stahelin RV, Kutateladze TG, Sansom MS (2011) Biophysical and computational studies of membrane penetration by the GRP1 pleckstrin homology domain. Structure 19:1338–1346.
31. Kalli AC, Devaney I, Sansom MS (2014) Interactions of phosphatase and tensin homologue (PTEN) proteins with phosphatidylinositol phosphates: insights from molecular dynamics simulations of PTEN and voltage sensitive phosphatase. Biochemistry 53:1724–1732.
32. Kalli AC, Sansom MS (2014) Interactions of peripheral proteins with model membranes as viewed by molecular dynamics simulations. Biochem Soc Trans 42:1418–1424.
33. Jeanmougin F, Thompson JD, Gouy M, Higgins DG, Gibson TJ (1998) Multiple sequence alignment with Clustal X. Trends Biochem Sci 23:403–405.
34. Tamura K, Stecher G, Peterson D, Filipski A, Kumar S (2013) MEGA6: Molecular Evolutionary Genetics Analysis version 6.0. Mol Biol Evol 30:2725–2729.
35. Jones DT (1999) Protein secondary structure prediction based on position-specific scoring matrices. J Mol Biol 292:195–202.
36. Cole C, Barber JD, Barton GJ (2008) The Jpred 3 secondary structure prediction server. Nucleic Acids Res 2008;36:W197–W201.
37. Yan R, Xu D, Yang J, Walker S, Zhang Y (2013) A comparative assessment and analysis of 20 representative sequence alignment methods for protein structure prediction. Sci Rep 3:2619.
38. Kozlowski LP, Bujnicki JM (2012) MetaDisorder: a meta-server for the prediction of intrinsic disorder in proteins. BMC Bioinform 2012;13:111.
39. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25: 3389–3402.
40. Rose PW, Prlic A, Bi C, Bluhm WF, Christie CH, Dutta S, Green RK, Goodsell DS, Westbrook JD, Woo J, Young J, Zardecki C, Berman HM, Bourne PE, Burley SK (2015) The RCSB Protein Data Bank: views of structural biology for basic and applied research and education. Nucleic Acids Res 43:D345–D356.
41. Sali A, Blundell TL (1993) Comparative protein modelling by satisfaction of spatial restraints. J Mol Biol 234:779–815.
42. Shen MY, Sali A (2006) Statistical potential for assessment and prediction of protein structures. Protein Sci 15:2507–2524.
43. Fiser A, Sali A (2003) ModLoop: automated modeling of loops in protein structures. Bioinformatics 19:2500–2501.
44. Xu D, Zhang Y (2012) *Ab initio* protein structure assembly using continuous structure fragments and optimized knowledge-based force field. Proteins 80:1715–1735.

45. Lovell SC, Davis IW, Arendall WB, III, de Bakker PI, Word JM, Prisant MG, Richardson JS, Richardson DC (2003) Structure validation by Calpha geometry: phi,psi and Cbeta deviation. Proteins 50: 437–450.
46. Sippl MJ (1993) Recognition of errors in three-dimensional structures of proteins. Proteins 17:355–362.
47. Wiederstein M, Sippl MJ (2007) ProSA-web: interactive web service for the recognition of errors in three-dimensional structures of proteins. Nucleic Acids Res 35:W407–W410.
48. Benkert P, Tosatto SC, Schomburg D (2008) QMEAN: a comprehensive scoring function for model quality assessment. Proteins 2008;71:261–277.
49. Jo S, Vargyas M, Vasko-Szedlar J, Roux B, Im W (2008) PBEQ-Solver for online visualization of electrostatic potential of biomolecules. Nucleic Acids Res 2008;36:W270–W275.
50. Van Der SD, Lindahl E, Hess B, Groenhof G, Mark AE, Berendsen HJ (2005) GROMACS: fast, flexible, and free. J Comput Chem 2005;26:1701–1718.
51. Abrahama MJ, Murtolad T, Schulzb R, Pálla S, Smith JC, Hessa B, Lindahla E (2016) GROMACS: high performance molecular simulations through multi-level parallelism from laptops to supercomputers. SoftwareX 1–2:19–25.
52. Hermans J, Berendsen HJC, van Gunsteren WF, Postma JPM (1984) A consistent empirical potential for water–protein interactions. Biopolymers 23:1513–1518.
53. Poger D, Mark AE (2010) On the validation of molecular dynamics simulations of saturated and *cis*-monounsaturated phosphatidylcholine lipid bilayers: a comparison with experiment. J Chem Theory Comput 2010;6:325–336.
54. Poger D, Van Gunsteren WF, Mark AE (2010) A new force field for simulating phosphatidylcholine bilayers. J Comput Chem 31:1117–1125.
55. Berger O, Edholm O, Jahnig F (1997) Molecular dynamics simulations of a fluid bilayer of dipalmitoylphosphatidylcholine at full hydration, constant pressure, and constant temperature. Biophys J 72:2002–2013.
56. Ejsing CS, Sampaio JL, Surendranath V, Duchoslav E, Ekroos K, Klemm RW, Simons K, Shevchenko A (2009) Global analysis of the yeast lipidome by quantitative shotgun mass spectrometry. Proc Natl Acad Sci USA 106:2136–2141.
57. Darden T, York D, Pedersen L (1993) Particle Mesh Ewald—an N. Log(N) method for Ewald sums in large systems. J Chem Phys 1993;98:10089–10092.
58. Essmann U, Perera L, Berkowitz ML, Darden T, Lee H (1995) A smooth particle Mesh Ewald method. J Chem Phys 1995;103:8577–8593.
59. Miyamoto S, Kollman PA (1992) Settle—an analytical version of the shake and rattle algorithm for rigid water models. J Comput Chem 13:952–962.
60. Hess B, Bekker H, Berendsen HJC, Fraaije JGEM (1997) LINCS: a linear constraint solver for molecular simulations. J Comput Chem 1997;18:1463–1472.
61. Bussi G, Donadio D, Parrinello M (2007) Canonical sampling through velocity rescaling. J Chem Phys 126:014101.
62. Nosé S, Klein ML (1983) Constant pressure molecular dynamics for molecular systems. Mol Phys 1983;50:1055–1076.
63. Parrinello M, Rahman A (1981) Polymorphic transitions in single crystals: a new molecular dynamics method. J Appl Phys 1981;52: 7182–7190.
64. Kabsch W, Sander C (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 22:2577–2637.
65. Kutzner C, Pall S, Fechner M, Esztermann A, de Groot BL, Grubmuller H (2015) Best bang for your buck: GPU nodes for GROMACS biomolecular simulations. J Comput Chem 36:1990–2008.
66. Rokas A, Holland PW (2000) Rare genomic changes as a tool for phylogenetics. Trends Ecol Evol 15:454–459.
67. Akiva E, Itzhaki Z, Margalit H (2008) Built-in loops allow versatility in domain-domain interactions: lessons from self-interacting domains. Proc Natl Acad Sci USA 2008;105:13292–13297.
68. Hashimoto K, Panchenko AR (2010)Mechanisms of protein oligomerization, the critical role of insertions and deletions in maintaining different oligomeric states. Proc Natl Acad Sci USA 2010;107: 20352–20357.
69. Brown JR, Auger KR (2011) Phylogenomics of phosphoinositide lipid kinases: perspectives on the evolution of second messenger signaling and drug discovery. BMC Evol Biol 11:4.
70. Baldauf SL, Palmer JD (1993) Animals and fungi are each other's closest relatives: congruent evidence from multiple proteins. Proc Natl Acad Sci USA 90:11558–11562.
71. Rivera MC, Lake JA (1992) Evidence that eukaryotes and eocyte prokaryotes are immediate relatives. Science 257:74–76.
72. Dunn B, Richter C, Kvitek DJ, Pugh T, Sherlock G (2012) Analysis of the *Saccharomyces cerevisiae* pan-genome reveals a pool of copy number variants distributed in diverse yeast strains from differing industrial environments. Genome Res 2012;22:908–924.
73. Chakravarty S, Godbole S, Zhang B, Berger S, Sanchez R (2008) Systematic analysis of the effect of multiple templates on the accuracy of comparative models of protein structure. BMC Struct Biol 8:31.
74. Peng J, Xu J (2011) A multiple-template approach to protein threading. Proteins 79:1930–1939.
75. Burden LM, Rao VD, Murray D, Ghirlando R, Doughman SD, Anderson RA, Hurley JH (1999) The flattened face of type II beta phosphatidylinositol phosphate kinase binds acidic phospholipid membranes. Biochemistry 38:15141–15149.
76. Divecha N (2010) Lipid kinases: charging PtdIns(4,5)P2 synthesis. Curr Biol 20:R154–R157.
77. Fairn GD, Grinstein S (2012) Cell biology. Precursor or charge supplier? Science 337:653–654.
78. Hadders MA, Williams RL (2010) Kinases charging to the membrane. Cell 143:865–867.
79. Hsing M, Cherkasov A (2008) Indel PDB: a database of structural insertions and deletions derived from sequence alignments of closely related proteins. BMC Bioinform 9:293.
80. Gupta RS, Nanda A, Khadka B (2017) Novel molecular, structural and evolutionary characteristics of the phosphoketolases from bifidobacteria and Coriobacteriales. PLoS One 12:e0172176.
81. Nandan D, Lopez M, Ban F, Huang M, Li Y, Reiner NE, Cherkasov A (2007) Indel-based targeting of essential proteins in human pathogens that have close host orthologue(s): discovery of selective inhibitors for *Leishmania donovani* elongation factor-1alpha. Proteins 67:53–64.
82. Cherkasov A, Nandan D, Reiner NE (2005) Selective targeting of indel-inferred differences in spatial structures of highly homologous proteins. Proteins 58:950–954.
83. Chan SK, Hsing M, Hormozdiari F, Cherkasov A (2007) Relationship between insertion/deletion (indel) frequency of proteins and essentiality. BMC Bioinform 2007;8:227.
84. Lopez M, Cherkasov A, Nandan D (2007) Molecular architecture of leishmania EF-1alpha reveals a novel site that may modulate protein translation: a possible target for drug development. Biochem. Biophys Res Commun 2007;356:886–892.
85. Moravcevic K, Mendrola JM, Schmitz KR, Wang YH, Slochower D, Janmey PA, Lemmon MA (2010) Kinase associated-1 domains drive MARK/PAR1 kinases to membrane targets by binding acidic phospholipids. Cell 143:966–977.
86. Wen W, Liu W, Yan J, Zhang M (2008) Structure basis and unconventional lipid membrane binding properties of the PH-C1 tandem of rho kinases. J Biol Chem 283:26263–26273.
87. Yan J, Wen W, Chan LN, Zhang M (2008) Split pleckstrin homology domain-mediated cytoplasmic-nuclear localization of PI3-kinase enhancer GTPase. J Mol Biol 378:425–435.

88. Kalli AC, Morgan G, Sansom MS (2013) Interactions of the auxilin-1 PTEN-like domain with model membranes result in nanoclustering of phosphatidyl inositol phosphates. Biophys J 2013;105:137–145.
89. Xu X, Song H, Qi J, Liu Y, Wang H, Su C, Shi Y, Gao GF (2016) Contribution of intertwined loop to membrane association revealed by Zika virus full-length NS1 structure. EMBO J 35:2170–2178.
90. Babakhani A, Gorfe AA, Gullingsrud J, Kim JE, Andrew MJ (2007) Peptide insertion, positioning, and stabilization in a membrane: insight from an all-atom molecular dynamics simulation. Biopolymers 85:490–497.
91. van der Rest ME, Kamminga AH, Nakano A, Anraku Y, Poolman B, Konings WN (1995) The plasma membrane of *Saccharomyces cerevisiae*: structure, function, and biogenesis. Microbiol Rev 1995;59: 304–322.
92. McGinnis MR, Tyring SK (2017) In Baron's medical microbiology, Albrecht T, Baron S, Castro G, Couch RB, Davis CP, Dianzani F, Mcginnis MR, Niesel DW, Woods GW editors, 4th ed. Galveston, Texas: University of Texas Medical Branch.
93. Fairn GD, McMaster CR (2005) Identification and assessment of the role of a nominal phospholipid binding region of ORP1S (oxysterol-binding-protein-related protein 1 short) in the regulation of vesicular transport. Biochem J 387:889–896.
94. Apsel B, Blair JA, Gonzalez B, Nazif TM, Feldman ME, Aizenstein B, Hoffman R, Williams RL, Shokat KM, Knight ZA (2008) Targeted polypharmacology: discovery of dual inhibitors of tyrosine and phosphoinositide kinases. Natl Chem Biol 4:691–699.
95. Engelman JA (2009) Targeting PI3K signalling in cancer: opportunities, challenges and limitations. Natl Rev Cancer 9:550–562.

CHAPTER 4

Novel molecular, structural and evolutionary characteristics of the Phosphoketolases from bifidobacteria and *Coriobacteriales*

This chapter describes the identification of multiple highly specific molecular markers in the forms of CSIs in phosphoketolase (PKs), a key enzyme involved in carbohydrate metabolism. The identified CSIs clearly distinguish the PKs of bifidobacteria from the phosphoketolase enzyme homologs found in most other bacteria. This chapter also highlights the evidence indicating the horizontal transfer of PKs gene between bifidobacteria and *Coriobacteriales* order of bacteria, which is comprised of saccharolytic organisms and also belongs to the phylum Actinobacteria. In addition, structural analyses and protein-protein docking study reveals the significance of some of the identified CSIs involved in the formation/stabilization of PKs dimer in bifidobacteria. My contribution towards the completion of this chapter includes data analysis, molecular modelling and structural analysis, protein-protein docking study, preparation and revision of the manuscript, and the production of main and supplemental figures and tables in the manuscript.

Due to limited space, supplementary materials (figures and tables) are not included in the chapter but can be accessed along with the rest of the manuscript at:

Gupta, R.S., Nanda, A., Khadka, B. (2017). Novel molecular, structural and evolutionary characteristics of the phosphoketolases from bifidobacteria and *Coriobacteriales*. *PloS One*, 12 (2), e0172176.

PLOS ONE

RESEARCH ARTICLE

Novel molecular, structural and evolutionary characteristics of the phosphoketolases from bifidobacteria and *Coriobacteriales*

Radhey S. Gupta*, Anish Nanda, Bijendra Khadka

Department of Biochemistry and Biomedical Sciences, McMaster University, Hamilton, Ontario, Canada

* gupta@mcmaster.ca

OPEN ACCESS

Citation: Gupta RS, Nanda A, Khadka B (2017) Novel molecular, structural and evolutionary characteristics of the phosphoketolases from bifidobacteria and *Coriobacteriales*. PLoS ONE 12 (2): e0172176. doi:10.1371/journal.pone.0172176

Editor: Eugene A. Permyakov, Russian Academy of Medical Sciences, RUSSIAN FEDERATION

Received: December 12, 2016

Accepted: January 12, 2017

Published: February 17, 2017

Data Availability Statement: All relevant data are within the paper and its Supporting Information files.

Funding: The work was supported by the research grant number 249924 from the Natural Sciences and Engineering Research Council of Canada. The funder had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

Abstract

Members from the order *Bifidobacteriales*, which include many species exhibiting health promoting effects, differ from all other organisms in using a unique pathway for carbohydrate metabolism, known as the "bifid shunt", which utilizes the enzyme phosphoketolase (PK) to carry out the phosphorolysis of both fructose-6-phosphate (F6P) and xylulose-5-phosphate (X5P). In contrast to bifidobacteria, the PKs found in other organisms (referred to XPK) are able to metabolize primarily X5P and show very little activity towards F6P. Presently, very little is known about the molecular or biochemical basis of the differences in the two forms of PKs. Comparative analyses of PK sequences from different organisms reported here have identified multiple high-specific sequence features in the forms of conserved signature inserts and deletions (CSIs) in the PK sequences that clearly distinguish the X5P/F6P phosphoketolases (XFPK) of bifidobacteria from the XPK homologs found in most other organisms. Interestingly, most of the molecular signatures that are specific for the XFPK from bifidobacteria are also shared by the PK homologs from the *Coriobacteriales* order of Actinobacteria. Similarly to the *Bifidobacteriales*, the order *Coriobacteriales* is also made up of commensal organisms, that are saccharolytic and able to metabolize wide variety of carbohydrates, producing lactate and other metabolites. Phylogenetic studies provide evidence that the XFPK from bifidobacteria are specifically related to those found in the *Coriobacteriales* and suggest that the gene for PK (XFPK) was horizontally transferred between these two groups. A number of the identified CSIs in the XFPK sequence, which serve to distinguish the XFPK homologs from XPK homologs, are located at the subunit interface in the structure of the XFPK dimer protein. The results of protein modelling and subunit docking studies indicate that these CSIs are involved in the formation/stabilization of the protein dimer. The significance of these observations regarding the differences in the activities of the XFPK and XPK homologs are discussed. Additionally, this work also discusses the significance of the XFPK-like homologs, similar to those found in bifidobacteria, in the order *Coriobacteriales*.

Introduction

Bifidobacteria are an important group of commensal microorganisms comprising significant constituents of the gastrointestinal tracts of humans, other mammals, as well as insects [1–4]. These bacteria are able to metabolize a wide variety of carbohydrates and glycans, and their genomes are particularly rich in carbohydrates-utilizing enzymes [5–10]. The saccharolytic ability of bifidobacteria plays a major role in their adaptation as important, and often dominant, inhabitants of the gut microbiota and possibly also for their health-promoting effects [1,5,6,9,11–13]. Bifidobacteria differ from other gut microbes in terms of their fermentation of carbohydrates. Specifically, these organisms lack the enzymes aldolase and glucose-6-phosphate (G6P) $NADP^+$ oxidoreductase, and as such, they are unable to utilize the conventional glycolysis pathway for carbohydrate metabolism [6]. Instead, bifidobacteria possess a unique fermentation pathway known as the "bifid shunt" which relies on the enzyme phosphoketolase to metabolize Fructose-6-phosphate (F6P) [3,14–19].

Phosphoketolases (PKs) are members of the thiamine pyrophosphate (TPP)-dependent enzyme family that play a crucial role in carbohydrate metabolism in various microbes [17,20]. Based on their substrate specificities, these enzymes can be categorized into two major groups. The common form of PK present in most organisms shows a substrate specificity mainly for xylulose-5-phosphate (X5P) (XPK, EC 4.1.2.9), and plays a fundamental role in pentose catabolism in obligatory and heterofermentative lactic acid bacteria (LAB), as well as in certain species of cyanobacteria and fungi via the phosphoketolase pathway [17,20–22]. In contrast, the form of PKs found in bifidobacteria (XFPK, EC 4.1.2.22) are unique in exhibiting comparable affinities for both X5P and F6P [16,18,23,24]. Thus, XFPK, in addition to splitting F6P into erythrose-4-phosphate and acetyl phosphate, also catalyzes the phosphorolysis of X5P into acetyl phosphate and D-glyceraldehyde-3-phosphate. Due to its ability to metabolize F6P, XFPK in bifidobacteria serves to link the carbohydrate metabolism pathway to the phosphoketolase pathway common to lactic acid group of bacteria [17,23,25–27]. Since the activity of PKs towards F6P is primarily found in bifidobacteria, the presence of the XFPK form of the enzyme is often used as a taxonomic tool for the identification of bifidobacteria [7,8,19,28,29].

Despite the well-known differences in the biological activities of PKs between bifidobacteria and other bacteria, very little is known about the molecular or biochemical basis accounting for the differences in the two forms of PKs. The amino acid sequences of XPK and XFPK exhibit more than 40% identity over their entire length, indicating that the two forms of enzymes are homologous [17,20,26]. The crystal structures of XFPK from *Bifidobacterium breve* and *Bifidobacterium longum*, with some bound cofactors and intermediates, have been solved [16,20]. The enzyme in bifidobacteria is a dimer with the active site located at the interface formed between the two subunits [16,20]. However, no structural information is available for the XPK form of the phosphoketolases.

Our recent comparative analyses of protein sequences from bifidobacteria have identified numerous conserved signature indels (CSIs) in various proteins involved in different cellular functions that are distinctive characteristics of the bifidobacterial homologs [30].

The present work focuses on the sequence features of the phosphoketolases to identify any characteristics that could prove helpful in understanding the differences between the two forms of PKs found in different organisms. These studies have led to the identification of multiple highly specific molecular differences in the forms of CSIs that clearly distinguish the XFPK of bifidobacteria from the XPK homologs found in most other bacteria. Interestingly, most of the molecular signatures that are specific for the XFPK from bifidobacteria are also shared by the PKs from the *Coriobacteriales* order of bacteria, which is comprised of saccharolytic organisms also belong to the phylum Actinobacteria [31,32]. Phylogenetic studies provide

evidence that the PKs in bifidobacteria are specifically related to those found in the *Coriobacteriales*, suggesting that the gene for PK (XFPK) was horizontally transferred between these two groups. The results of protein modelling and *in silico* docking studies presented here reveal that a number of the identified CSIs in the PK sequences which are distinguishing characteristics of the *Bifidobacteriales/Coriobacteriales* homologs are present at the subunit interface in the XFPK structure and they are involved in the formation/stabilization of the protein dimer. The significance of these observations regarding the differences in the activities of the XFPK and XPK are discussed.

Methods

Identification of conserved indels and phylogenetic tree construction

Conserved signature indels (i.e. insertions or deletions) in the sequence alignment of phosphoketolase were identified as described recently [30,33]. A multiple sequence alignment of the PK homologs from representative bifidobacteria and other bacterial phyla was created using ClustalX [34]. The alignment was visually inspected for the presence of different indels that were flanked on both sides by at least 5–6 conserved amino acid residues in the neighbouring 30–40 amino acids. For all indels meeting these criteria, detailed BLASTp searches were carried out on short sequence segments containing the indel and the flanking conserved regions (60–100 amino acids long) to determine the specificity of the indels. SIG_CREATE and SIG_STYLE (available on Gleans.net) were used to create signature files that are shown here [30,33]. Due to space limitations, sequence information for all *Bifidobacterium* or *Coriobacteriales* species (or subspecies) is not shown in the alignment files. However, unless otherwise noted, all of the described CSIs are specific for the indicated groups (i.e. similar CSIs were not present in the protein homologs from other bacteria in the top 500 Blast hits) [30]. For phylogenetic analysis, a multiple sequence alignment of PK homologs was constructed from bifidobacteria, *Coriobacteriales* and a limited number of outgroup species (Firmicutes and Actinobacteria). After removing areas of poor sequence conservation using the Gblocks 0.91b program [35], a maximum likelihood (ML) tree based on the resulting alignment was constructed using the MEGA 6 program [36] employing the Jones-Taylor-Thornton [37] and Whelan and Goldman [38] substitution models, respectively.

Structural analysis of the CSIs and homology modeling of phosphoketolase homologs

The structural models of the PK (or XFPK) protein from several bifidobacteria species (viz. *B. breve*, *B. bifidum*, *B. animalis*, *B. reuteri*) were generated using homology modelling. The secondary structure analyses on the selected homolog sequences were initially performed via PSIPRED v3.3 web server [39].The crystalized *B. longum* PK structure (PDB ID: 3AI7) was utilized as a template and the comparative modeling was carried out using MODELLER v9.11 [40]. Initially, 200 models were generated and ranked/selected using Discrete Optimized Protein Energy (DOPE) scores [41]. The secondary structure elements in the regions containing CSIs were examined and compared with results of the PSIPRED analysis to ensure their reliability. The stereo-chemical properties of the final models were assessed using three independent servers: RAMPAGE, ERRAT, and Verify3D [42], [43] [44,45]. These tools use a dataset of highly refined structures to evaluate the statistical significance of models based on the conformation, location, and the environment of each amino acid in the sequence, as well as the model's overall structural stability. The superimposition of the validated models with the template

structures was carried out using PyMOL (Version 1.7.4; Schrödinger, LLC.) to examine the structure and location of identified CSIs in the PK (or XFPK) structures..

Identification of the macromolecular interface formed between the individual subunit and the residues in the CSIs that are involved in subunit-subunit interactions was determined by submitting the three-dimensional coordinate file of the *B. longum* PFK dimeric structure to the PDBePISA server using default parameters (Version 1.48)[46].

Protein-protein docking to examine the dimerization potential of the bifidobacteria PKs

The protein-protein docking approach was utilized to gain insights concerning the structural roles of the interface interacting residues and to study the dimerization potentials of the PK homologs from *Bifidobacterium* species. The structural models of the CSI-lacking forms of PKs from a number of bifidobacterial species (viz. *B. longum*, *B. breve*, *B. bifidum*, *B. animalis* and *B. reuteri*) were generated using homology modelling methods as described above. The known structures and the individual structural models of PFK monomers forms were submitted to three fully automated web-based protein-protein docking programs, viz. ZDOCK (Version 3.0.2)[47], PatchDock (Version Beta 1.3) [48], and ClusPro 2.0 [49] using default parameters. ZDOCK utilizes grid-based fast Fourier transform (FTT) for efficient global search of docking orientation between two proteins [47]. Its scoring function is based on pairwise shape complementarity, electrostatics, and a pairwise atomic statistical potential developed using contact propensities of transient protein complexes. PatchDock is a very efficient geometry-based molecular docking algorithm which is aimed to yield the good molecular shape complementarity of protein-protein complexes [48]. Its scoring function includes both geometric fit and atomic desolvation energy [48]. ClusPro utilizes PIPER, a rigid body docking program [50], which is based on a novel FFT based docking approach with pairwise potential [49,50]. The structures with maximum cluster size and the conformation closest to the solved crystal structure of PFK dimer with lowest root mean square deviation (RMSD) was selected as a representative structure for the detailed interface interaction analysis. Visualization and structure alignment of the CSI-containing and CSI-lacking dimer structure was carried out using PyMOL (Version 1.7.4; Schrödinger, LLC.). PDBePISA (Version 1.48)[46] server was used for detailed interface analysis.

Results

Distinguishing features of the phosphoketolase sequences from *Bifidobacteriales* and *Coriobacteriales*

The PKs from bifidobacteria differ from other studied bacteria because of their ability to utilize/metabolize both F6P and X5P. To gain insights into the molecular basis of the differences in the biochemical properties of PKs from bifidobacteria (XFPK) versus other bacteria (XPK), a multiple sequence alignment of representative PK homologs from different bacterial groups was constructed. Examination of this sequence alignment has identified a number of conserved indels that are uniquely present in the PK homologs from bifidobacteria as well as those from members of the order *Coriobacteriales*, but not found in the homologs from other groups/phyla of bacteria. In Fig 1, we present excerpts from sequence alignment of PKs showing a number of conserved signatures indels (CSIs) found in this protein that are distinctive characteristics of the *Bifiodobacteriales/Coriobacteriales* homologs. The first of these CSIs (CSI #1(11)) is a 3 aa insertion that is commonly shared by all *Bifidiobacteriales* homologs (all available homologs without any exception) as well as different *Coriobacteriales* homologs. Within

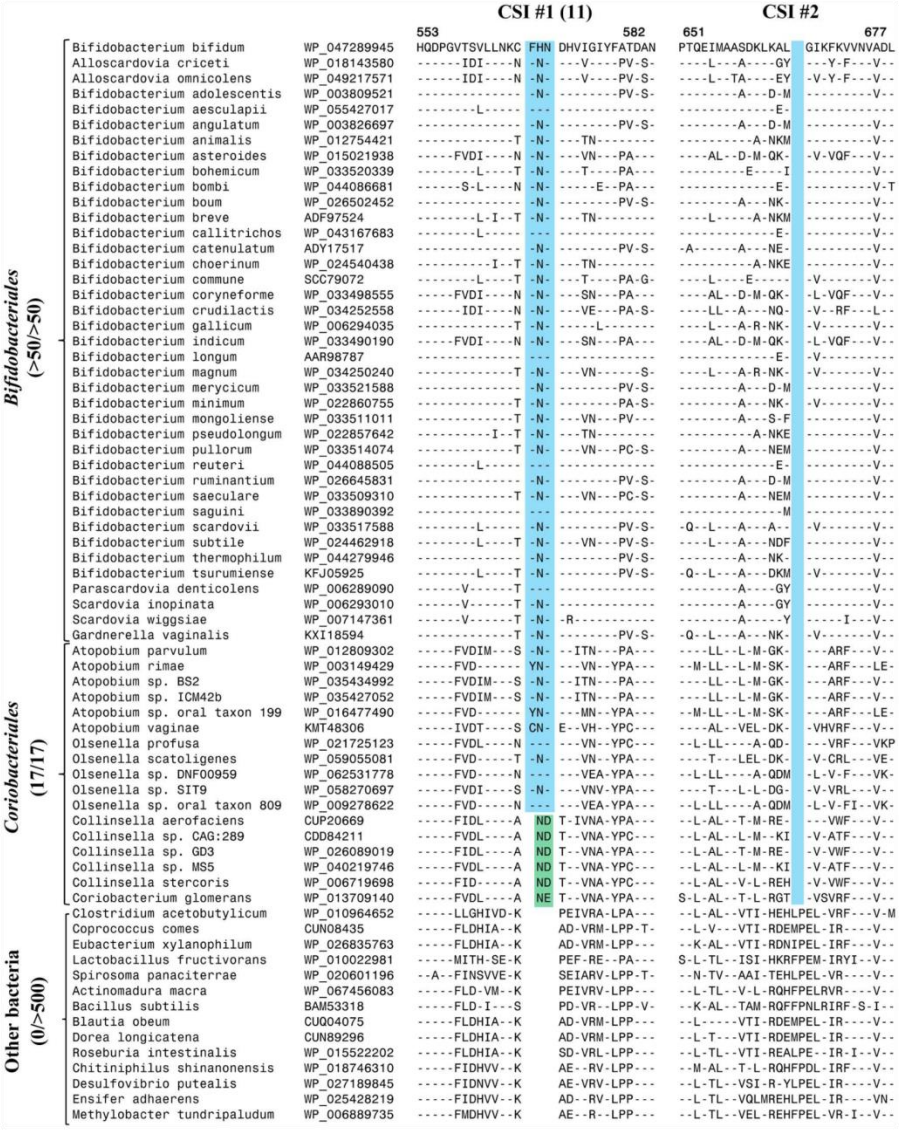


Fig 1. Excerpts from a sequence alignment of phosphoketolases showing a number of conserved signature indels (CSIs) that are either uniquely found in members of the orders *Bifidobacteriales* and *Coriobacteriales* or are commonly shared by the members of these two orders. For the CSI # 1(11) shown in this figure, CSI #1 refers to the 3 aa insert that is specific for the bifidobacteria and most *Coriobacteriales*, whereas the CSI # 11 corresponds to the 2 aa insert present in the same position in members of the genera *Collinsellla* and *Coriobacterium*. The dashes (-) in this alignment as well as in all other alignment figures indicate identity with the amino acid on the top line. Sequence information is presented for only a limited number of species. However, unless otherwise indicated the described CSIs are specific for the indicated groups of bacteria.

doi:10.1371/journal.pone.0172176.g001

the *Coriobacteriales*, while the *Atopobium* and *Olsenella* spp. contain a 3 aa insertion similar to that found in the bifidobacteria, a shorter 2 aa insert is present in the *Collinsella* and *Coriobacterium* spp. (the shorter insert found in the latter taxa at the same position is referred to as CSI #11). The second CSI shown on the right hand side in Fig 1 (CSI # 2) is comprised of a 2 aa deletion that is specifically found in all *Bifidobacteriales* and *Coriobacteriales* homologs. Both of the identified CSIs are flanked by conserved regions and, except for their shared presence in all sequenced *Bifidobacteriales/Coriobacteriales* homologs, they are not found in homologs from any other bacteria (within the top 500 Blast hits). In addition to the CSIs shown in Fig 1, 4 other CSIs were identified in the sequence alignments of PKs (CSIs # 3–6), which are also commonly shared by all PK homologs from the *Bifidobacteriales/Coriobacteriales*. Sequence information for these CSIs (#3–6) is provided in Figures A, B and C in S1 Fig. These CSIs are also either uniquely or mainly found in the PK homologs from *Bifidobacteriales* and *Coriobacteriales*. However, in some of these cases, CSIs of similar lengths are also present in a limited number (≈ 5%) of other unrelated bacteria.

The sequence alignment of PK sequences also contains a number of additional CSIs where inserts of different lengths are present in the same positions in the *Bifidobacteriales* and *Coriobacteriales* homologs. These CSIs permit differentiation among the PK homologs found in these two orders of bacteria, and also between certain members of these two orders. In Fig 2A, sequence information is presented for a conserved region where a 2 aa insert is present in all of the PK homologs from bifidobacteria (CSI #7), whereas the homologs from *Coriobacteriales* contain a 3 aa long insertion (CSI #9) in the same position. In another location within the sequence alignment of PKs (Fig 2B), a 3 aa insert is present in the PK homologs of most bifidobacteria (CSI #8), whereas the homologs from *Coriobacteriales*, as well as a number of deep-branching members of the order *Bifidobacteriales*, were found to contain a 2 aa insertion in the same position (CSI #10). Lastly, one additional large CSI (11 aa long insertion) present in the PK homologs is only found in the *Coribacteriales* homologs belonging to the genera *Collinsella*, *Coriobacterium* and *Olsenella*, but it is not found in members of the genus *Atopobium* or *Bifidobacteriales* species. Sequence information for this CSI (CSI #12), is presented in Figure D in S1 Fig.

Phylogenetic branching pattern of the PKs indicate horizontal gene transfer from *Coriobacteriales* to the *Bifidobacteriales*

The shared presence of multiple CSIs by the PFK homologs from bifidobacteria and *Coriobacteriales* strongly suggests that the homologs from these two groups are closely related. Although the orders *Bifidobacteriales* and *Coriobacteriales* are both part of the phylum Actinobacteria, in phylogenetic trees based on 16S rRNA and other genes/proteins sequences, members of these two orders exhibit distinct branching [51–54]. In contrast to the *Bifidobacteriales*, which branch in the proximity of the order *Actinomycetales*, the *Coriobacteriales* species, along with the other members of the class *Coriobacteriia*, form one of the deepest branching lineages within Actinobacteria [51,53,54]. To understand the significance of the shared presence of multiple highly-specific sequence features by these two groups of bacteria, a phylogenetic tree based on the sequences of PK homologs was constructed. The maximum-likelihood tree based on PK sequences, shown in Fig 3, contains information for all bifidobacteria and *Coriobacteriales* homologs as well as limited representatives from other orders of Actinobacteria, and also some sequences from the deeper branching Firmicutes phylum. In this tree, which was rooted using sequences from the Firmicutes species, the homologs from bifidobacteria and *Coriobacteriales* formed a strongly supported clade, which branched deeply in comparison to the homologs from other actinobacteria, and this clade was separated from all other bacteria by a long

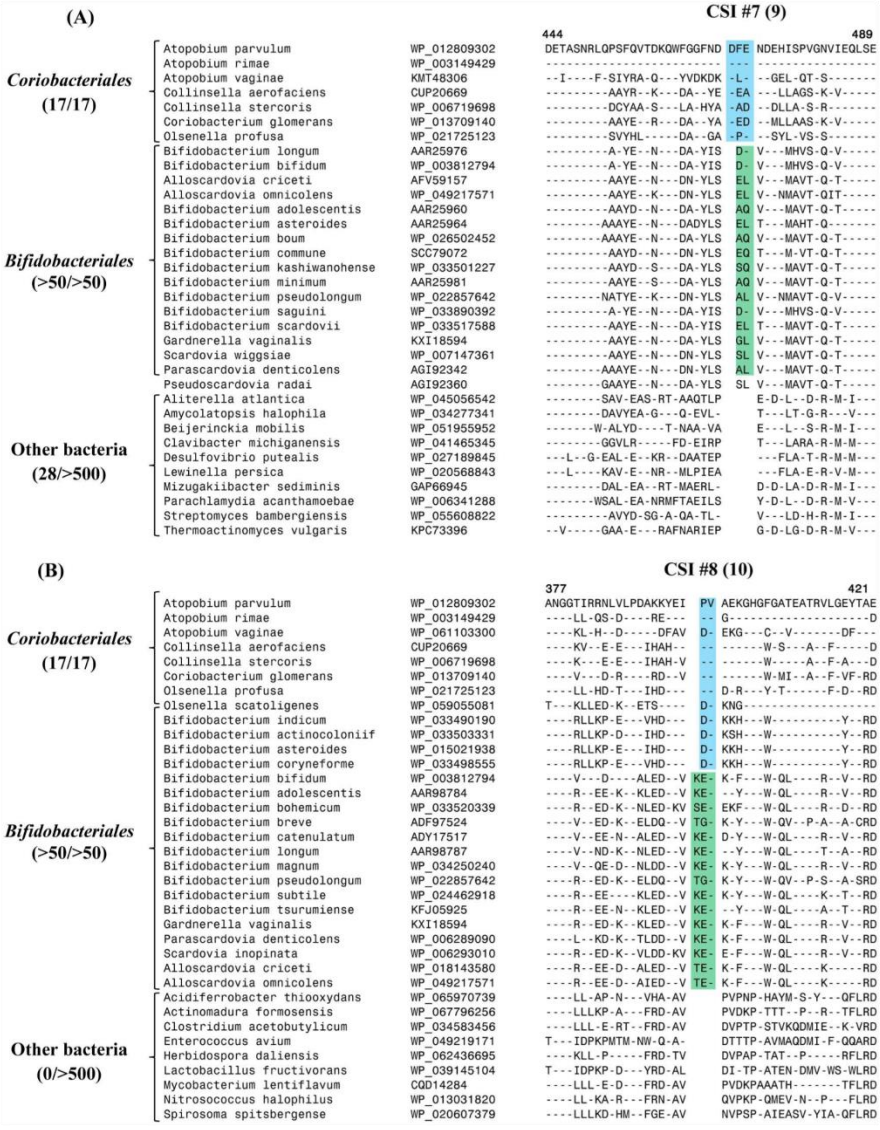


Fig 2. Partial sequence alignments of phosphoketolases showing a number of conserved signature indels (CSIs) where indels of different lengths are present in the same positions in members of the orders *Bifidobacteriales* and *Coriobacteriales*. When two different CSIs are present in the same position, in our numbering scheme, the first number refers to the CSI found in *Bifidobacteriales* group, whereas the second number in parenthesis describes the CSI found in the *Coriobacteriales*. Thus, CSI #7 and CSI #8 shown in this figure (parts A and B) refer to the indels present in all or most *Bifidobacteriales*, whereas CSI #9 and CSI #10 describe the indels found either only in the *Coriobacteriales* or in the *Coriobacteriales* plus certain deep branching bifidobacteria. The dashes (-) in the alignment indicate identity with the amino acid on the top line. The evolutionary interpretation of these indels is provided in the text and in Fig 3.

doi:10.1371/journal.pone.0172176.g002

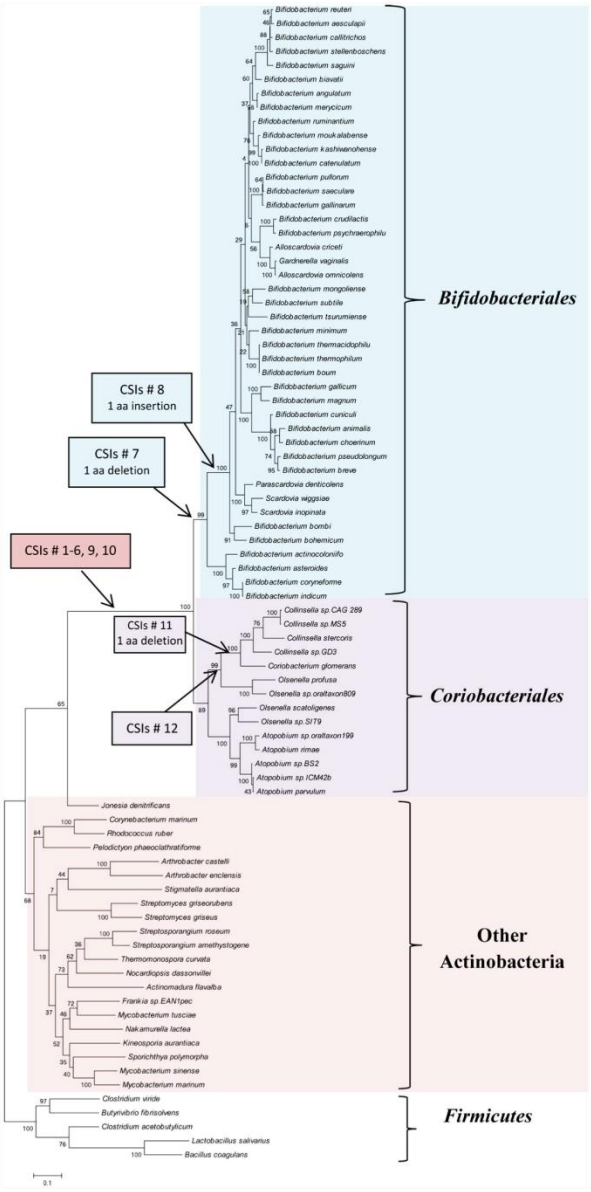


Fig 3. A maximum likelihood distance tree based on PKs sequences for members of the phylum Actinobacteria are representative outgroup species from the phylum *Firmicutes*. The numbers on the nodes indicate bootstrap scores for the group of species represented by different nodes. In this tree, members of the order *Bifidobacteriales* branch with the *Coriobacteriales* and the clade comprising of these two orders is separated from all other Actinobacteria/bacteria by a long branch. Based on the species distributions of different CSIs, the evolutionary stages where genetic changes giving rise to different CSIs have likely occurred are marked.

doi:10.1371/journal.pone.0172176.g003

branch. The observed branching and the strong affinity of the bifidobacterial homologs with the *Coriobacteriales* in the PK tree is in contrast to the distinct branching of the members of these two orders in the 16S rRNA tree and phylogenetic trees based on other genes/proteins sequences [51,53,54]. The observed results strongly suggests that the gene for the PK has been horizontally transferred between these two orders of Actinobacteria, and based on the known deeper branching of the order *Coriobacteriales* [31,51,53], the gene transfer has likely occurred from a *Coriobacteriales* to the *Bifidobacteriales*.

The inference from phylogenetic studies that the PK gene in *Bifidobacteriales* has been acquired from *Coriobacteriales* permits us to offer the most parsimonious explanation for the species distribution of different CSIs that are found in the PK homologs from these two groups of bacteria. Thus, the CSIs #1–6, where CSIs of similar lengths are present in different *Coriobacteriales* and *Bifidobacteriales* homologs, were likely present in the transferred *Coriobacteriales* PK gene. It can also now be inferred that the transferred *Coriobacteriales* PK gene contained a 3 aa insertion where the CSI #7(9) is found and an insertion of 2 aa, where the CSI #8(10) is present (Fig 2). Subsequent to the acquisition of this PK gene by a common ancestor of the *Bifidobacteriales*, further changes have occurred in this region that account for the differences in the lengths of the CSI #7 versus CSI #9, and CSI #8 versus CSI #10 between the *Coriobacteriales* and the *Bifidobacteriales*. These changes include a 1 aa deletion in the PK gene in the common ancestor of bifidobacteria where the CSI #7 is found, and a 1 aa insertion in the PK gene in the common ancestor of bifidobacteria, except the deepest branching members, where the CSI # 8 is found. The species specificities of different identified CSIs and the evolutionary stages where the genetic changes which gave rise to these CSIs have likely occurred are marked in the phylogenetic tree shown in Fig 3.

The species distributions of different CSIs also provide insights concerning the *Coriobacteriales* taxa from which the PK gene was likely transferred to the *Bifidobacteriales*. Of the described CSIs, the CSI #12 (Figure D in S1 Fig) is a specific characteristic of the PK homologs belonging to the genera *Collinsella*, *Coriobacterium* and *Olsenella*, but it is lacking in members of the genus *Atopobium*. The absence of this large CSI in all *Bifidobacteriales* homologs provides evidence that the PK gene was not acquired from members of the order *Coriobacteriales* which contain this CSI, but instead it originated from a member of this order lacking this CSI, such as a members of the genus *Atopobium* or closely related taxa. The CSI #1(11) (Fig 1) where a 3 aa insertion is found in all *Coriobacteriales* and *Bifidobacteriales* PK homologs, except those from the genera *Collinsella* and *Coriobacterium* also provides evidence that the transferred PK gene was not derived from these two genera of the *Coriobacteriales*.

Locations of the CSIs in the phosphoketolase structure and their possible significance

The phosphoketolase protein is comprised of three domains: N-terminal PP-domain (PP-D), middle PYR-domain (PYR-D), and the C-terminal domain (CT-D). The locations of the different identified CSIs in the primary structure of the *B. longum* protein are depicted in Fig 4. The insertions in the PK sequence in this figure are indicated by red-colored bold and underlined residues, whereas the deletions are present in between the residues marked in blue. Secondary structure elements of the sequence are displayed above the primary sequence, with helices shown as cylinders and sheets shown as arrows. As seen, the identified CSIs are present in different domains of the bifidobacteria PK homologs.

We have also mapped the locations of different CSIs in the crystal structure of PK from *B. longum* and a surface representation of the CSIs in the structure of a PK monomer (PDB ID: 3AI7) is shown in Fig 5. The three different domains in the protein are shown in three different

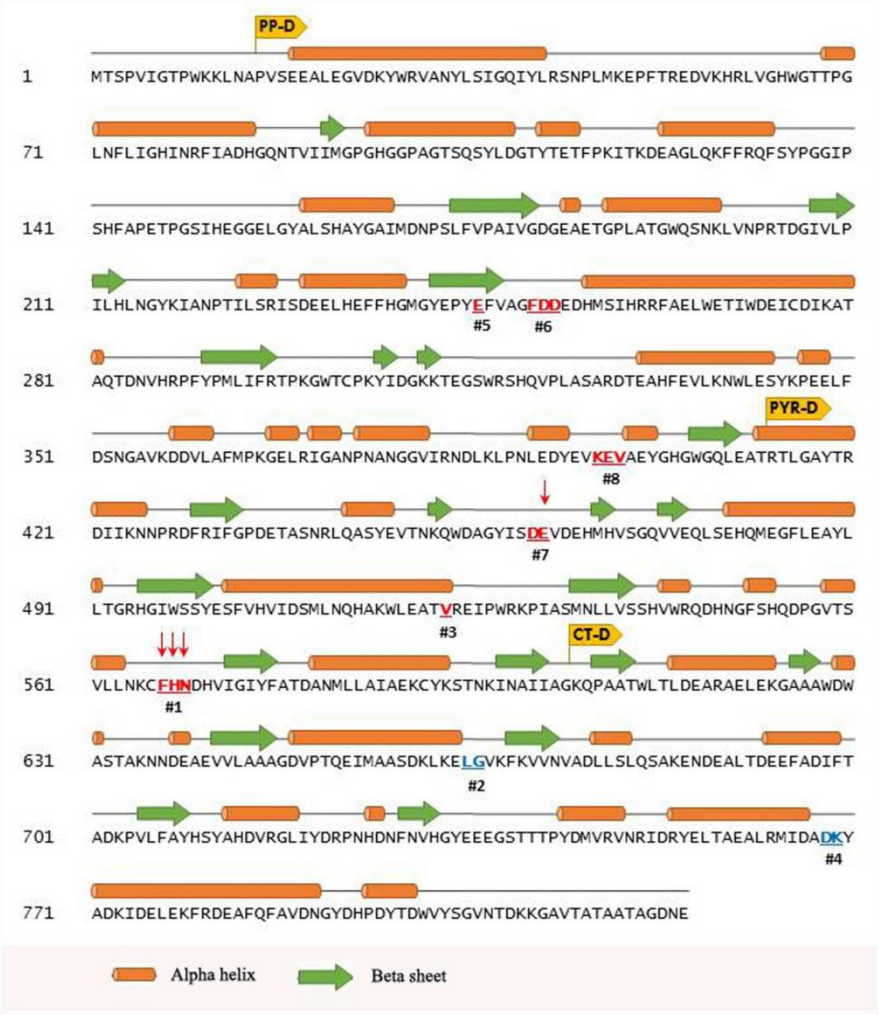


Fig 4. Primary sequence of phosphoketolase protein from *Bifidobacterium longum* depicting the location of different CSIs. Secondary structure elements are displayed above the primary sequence, with helices shown as cylinders and sheets shown as arrows. The secondary structure information was obtained directly from the solved structure of *B. longum* PK from Protein Data Bank (PDB ID: 3AI7). NCBI CD-search webserver was used to demarcate the three domains (yellow boxes): N-terminal PP-domain (PP-D), middle PYR-domain (PYR-D), and the C-terminal domain (CT-D). The eight identified CSIs in bifidobacteria PKs are indicated by bold and underlined residues, insertions are shown in red and the positions where deletions are present are shown in blue. The arrows on the top of CSIs indicate the residues contributing in subunit-subunit interactions, as determined via the PISA webserver.

doi:10.1371/journal.pone.0172176.g004

shades of green color; pale green as the N-terminal (PP) domain, lime green as the middle (PYR) domain, and forest green as the C-terminal domain. The bound TPP cofactor located at the active site is shown in magenta. Close-up views of the regions of the PK protein containing the eight bifidobacteria CSIs are shown in cartoon representation with the insertions depicted

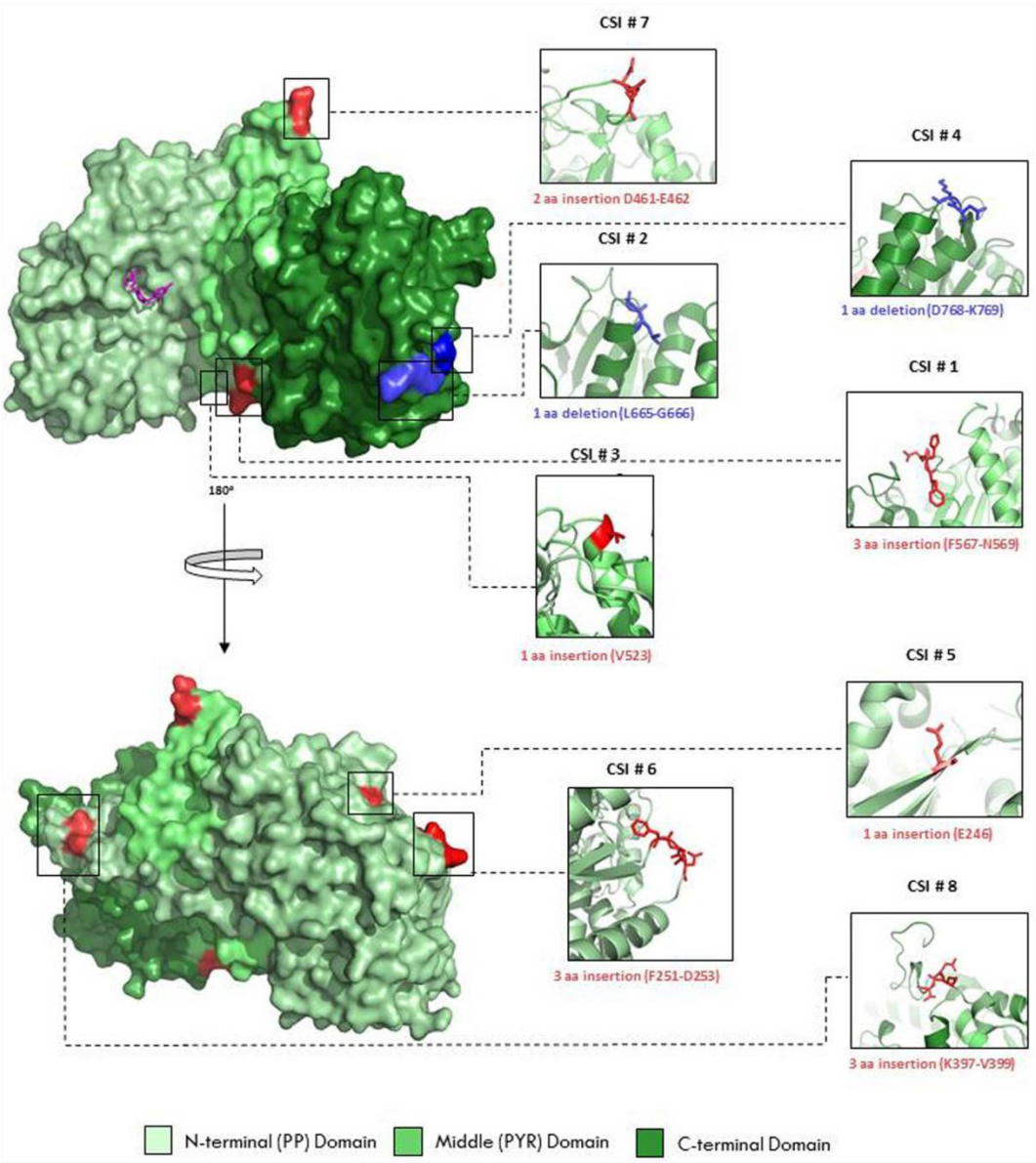


Fig 5. Surface representation of the phosphoketolase crystal structure monomer from *Bifidobacterium longum* (PDB ID: 3AI7). The three different domains in the protein are shown in three different shades of green color; pale green as N-terminal (PP) domain, lime green as Middle (PYR) domain, and forest green as C-terminal domain. The bound TPP cofactor located at the active site is shown in magenta. Close-up views of the regions containing eight different Indels are shown in cartoon representation with the insertions are depicted as red and the deletions as blue. The locations of the two deletions are highlighted by coloring the flanking residues as blue.

doi:10.1371/journal.pone.0172176.g005

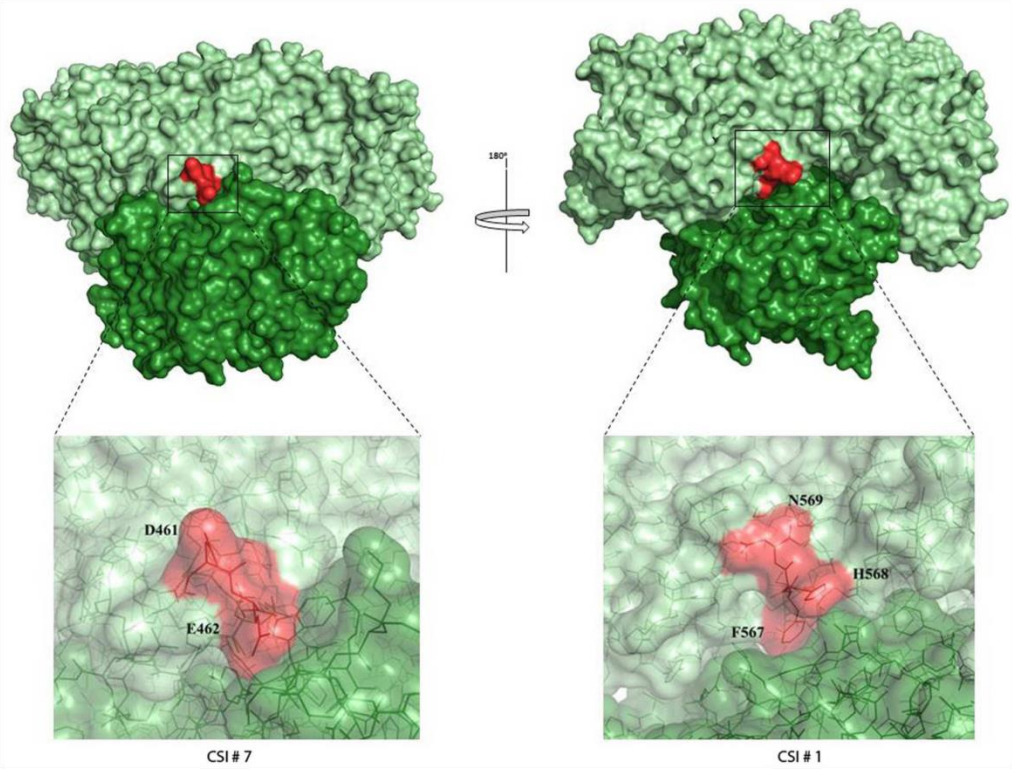


Fig 6. Surface representation of the phosphoketolase crystal structure dimer from *Bifidobacterium longum* (PDB ID: 3AI7). Individual monomers are shown in two different shades of green (pale green and forest green). The residues from two different CSIs (#1 and # 7) that are indicated to be involved in the dimer formation are highlighted red. The lower figures show close-up views of the 2 aa (CSI # 7) (left; D461-E462) and the 3aa insertion (CSI # 1) (right, F567-N569) (shown in red). Individual residues are labelled and clearly show their close proximity to the other subunit.

doi:10.1371/journal.pone.0172176.g006

in red and the deletions in blue (Fig 5). As seen from Figs 4 and 5, most of the identified CSIs in the PK protein are present in between the secondary structure elements found in the protein and most of them are located on the surface of the PK monomer. The only exceptions seen are CSIs #3 and #5, where single amino acid insertions have occurred at the end of a helix or beta sheet leading to possible lengthening of these structural elements.

The functional PK enzyme in bifidobacteria is a dimer with the active site located between the subunit interface. To explore the macromolecular interface formed between the individual monomers and to determine if any of the CSIs in the bifidobacteria PK are involved in the interaction between the subunits, the structural coordinate file for one of the PK subunit from *B. longum* (PDB ID: 3AI7) was submitted to the PDBePISA server (Version 1.48; Krissinel and Henrick, 2007). A surface representation of the phosphoketolase dimer from *B. longum* showing the subunit interaction is shown in Fig 6. Individual monomers in this figure are shown in two different shades of green (pale green and forest green). As seen from Fig 6A, the residues from two different CSIs (shown in red) are located at the subunit interface and are indicated to

Table 1. Protein-protein docking results of *Bifidobacterium* XFPK structure models for the CSIs-containing and CSIs-lacking proteins.

	CSI-containing Structure			CSI-lacking Structure		
Homolog	**ZD**	**PD**	**CP**	**ZD**	**PD**	**CP**
Bifidobacterium bifidum	5,565.215	39,902	-3,141.1	1,331.985	39,648	-2,301.8
Bifidobacterium animalis	3,206.406	54,118	-2,586.3	2,195.55	25,188	-,2,355
Bifidobacterium longum	3,217.031	48,084	-3,673.6	2,044.969	28,778	-2,572.4
Bifidobacterium reuteri	4947.682	51,414	-2,820.4	2,974.984	45,922	-2,609.8
Bifidobacterium breve	6,422.722	26, 438	-3,446.4	2,114.679	20,230	-2,591.3

Three different servers viz. Z-DOCK score (ZD), PatchDock score (PD), and ClusPro (CP) were utilized to create the dimer complex of PFK. The docking results are shown as Z-DOCK score, PatchDock geometry shape complementary score, and ClusPro (CP) lowest energy score (negative value). The removal of the CSIs # 1 and #7 from PFK homolog structures from *Bifidobacterium* species resulted in a dimer with decreased docking score when compared to the docking scores of unmodified (CSI-containing) proteins, as determined by all three servers.

doi:10.1371/journal.pone.0172176.t001

be involved in dimer formation. Close-up views of the two CSIs which are located at the subunit interface, viz. CSI #7 (D461-E462) and CSI #1 (F567-N569), are shown in Fig 6B. Of the CSIs located at the interface, several residues are involved in specific interactions; GLU (E) at position 462 is involved in hydrogen bonding, PHE (F) at position 567 is an interface residue, and HIS (H) at position 568 is involved in salt bridge formation.

To explore the roles of the CSIs #1 and #7 in the formation/stabilization of the PK dimers in bifidobacteria, dimerization potentials of the bifidobacterial PK homologs with and without these CSIs were investigated by means of protein-protein docking studies. For these studies, both the known structures, as well as the validated homology models of PK from several bifidobacteria species which either contained or lacked the CSIs # 1 and #7, were submitted to the three online protein docking servers: ZDOCK (Version 3.0.2; Pierce et al., 2011), PatchDock (Version Beta 1.3; Duhovny et al., 2002), and ClusPro 2.0 (Comeau et al., 2004). The docking scores obtained from the three different servers are shown in the Table 1. As seen from Table 1, the docking scores (i.e. dimerization potentials) of PFK homologs that contained the CSIs were much higher in comparison to those obtained with the CSI-lacking homologs, and all three docking servers yielded similar results. The results from the ZDOCK-server, which consistently produced dimer conformations with lower RMSD values compared to the other servers, were then uploaded to PDBePISA (Version 1.48; Krissinel and Henrick, 2007) for detailed interface analysis. The representative structure of *B. breve* PFK dimer structure containing CSIs as well as lacking the CSI #1 and CSI #7, obtained from ZDOCK is shown in the Fig 7. As shown in Fig 7, the residues in the CSI #1 and CSI #7 (labelled and highlighted red) provide additional surface area for binding and interaction at the interface and for dimer formation. The removal of the residues corresponding to these CSIs in the region resulted in the loss of an interacting surface at the interface (Fig 7c).

Discussion

Bifidobacteria differ from all other microbes in using a unique fermentation pathway known as the "bifid shunt" for the metabolism of different carbohydrates [5–7,14]. A key component of the "bifid shunt" enabling carbohydrate metabolism via this pathway is the presence of a novel form of the enzyme phosphoketolase (XFPK), which, in addition to carrying out phosphorolysis of X5P, is also able to convert F6P into erythrose-4-phosphate and acetyl phosphate [18,20,24,29]. The existence of the bifid shunt allows bifidobacteria to produce more ATP from carbohydrates than through other conventional pathways [5,6,20]. Specifically, the bifid shunt yields 2.5 ATP per mole of glucose compared to 2 ATP per glucose formed via the

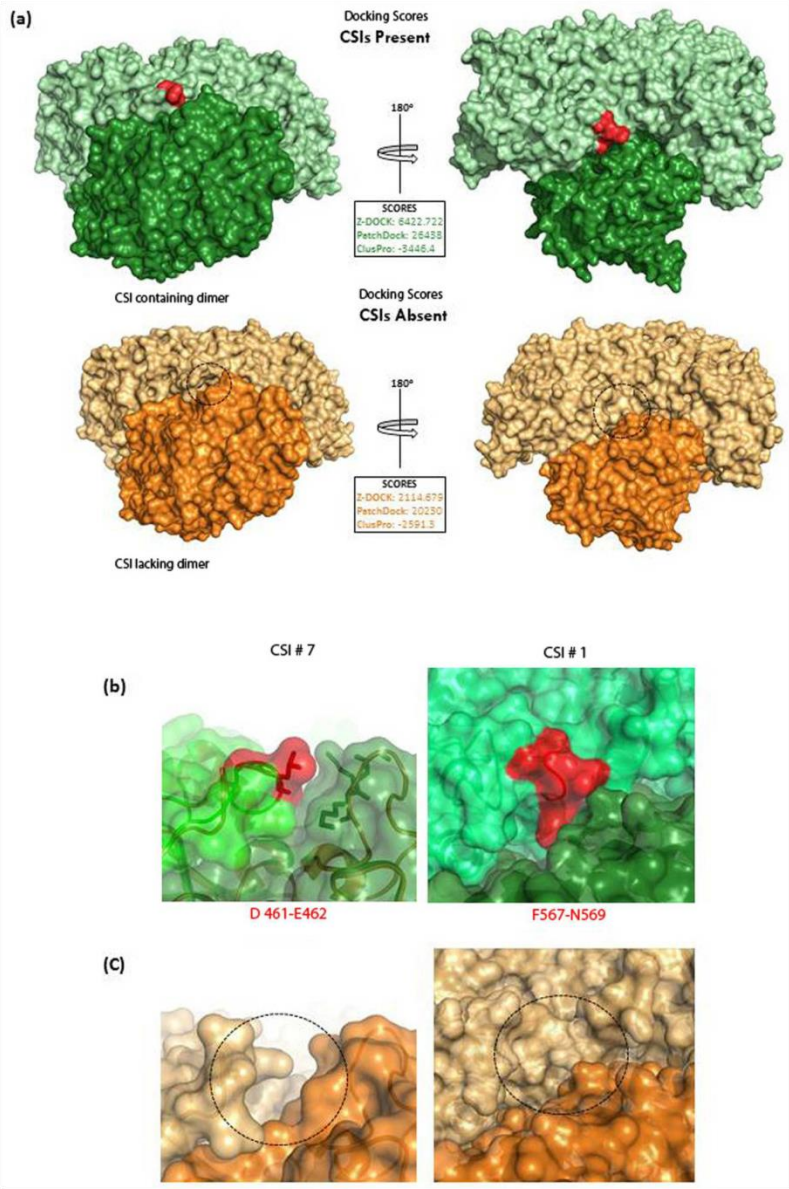


Fig 7. (a) The crystal structure of *Bifidobacterium breve* phosphoketolase dimer (PDB ID: 3AHC; green) with all CSIs present in the structure (red) versus model structure of the *B. breve* phosphoketolase dimer with the CSIs #1 and # 7 removed (orange). The dimers were generated by submitting the individual monomers (crystal structures and homology models) to three webservers: Z-DOCK, PatchDock, and ClusPro. The difference in docking scores generated by these servers is shown. The representative dimer PK from *B. berve* was calculated using ZDOCk. Removal of the CSIs resulted in small

cavities along the dimer interface (indicated by dashed circle) and a significantly reduced score. **(b)** Close-up views of the residues from 2 aa insertion and 3 aa insertion show that these CSIs are located at the dimer interface and directly involved in interactions with the other subunit (indicated by an arrow). **(c)** Close-up views of the region in the protein from where the residues corresponding to 2 aa insertion (CSI # 7) and 3 aa insertion (CSI # 1) were removed. The gaps created by the removal of these CSI are indicated by dashed circle.

doi:10.1371/journal.pone.0172176.g007

Embden-Meyerhof-Parnas glycolytic pathway [5,6,20]. Each mole of glucose also leads to the formation of 1.5 moles of acetate and 1 mole of lactate. The formation of these metabolites is of great benefit to the host organisms; the acetate produced in the gut is transported to the liver and used for the production of ATP, whereas lactate possesses anti-microbial activity and prevents proliferation of potential pathogens [5,5,6,20,55,56].

Phosphoketolases exhibiting high degree of sequence similarity to the bifidobacterial XFPK are widely distributed among prokaryotic organisms, and certain eukaryotes, but they exhibit specificity for only X5P and are unable to metabolize F6P [17,22,26,57–59]. However, the molecular and/or structural characteristics that differentiate the XFPK from XPK, which may be responsible for the important differences in their biochemical properties, are not known at present. Analyses of PK sequences from different organisms carried out in this work have provided important insights in this regard. Based on comparative analyses of PK sequences, this work has identified multiple high-specific sequence features in the forms of CSIs in the PK sequences that clearly distinguish the XFPK homologs of bifidobacteria from the XPK homologs found in most other organisms. An interesting and unexpected result is the discovery that the XFPK homologs from bifidobacteria are closely related to those found in the order *Coriobacteriales* and that most of the CSIs that are distinctive characteristics of the bifidobacteria XFPK are also present in the *Coriobacteriales* PKs. Phylogenetic studies on PK sequences show that the homologs from bifidobacteria from a strongly supported clade with the *Coribacteriales* PKs and the observed branching pattern of species from these two orders is different than that seen in phylogenetic trees based on other gene/protein sequences [51,53,54]. The observed branching pattern strongly suggests a horizontal transfer of the PK gene between these two orders of Actinobacteria. The phylogenetic branching pattern and the species distribution of different identified CSIs suggest that the PK gene was horizontally transferred from a *Coriobacteriales* to the common ancestor of the *Bifidobacteriales* that the *Coriobacteriales* taxon from which the PK gene was acquired likely corresponded to a member of the genus *Atopobium* or a closely related species.

The findings from this study indicate that XFPK homologs from bifidobacteria differ from all other XPK homologs (except those from the *Coriobacteriales*) by many highly-conserved sequence features and they strongly suggests that the described sequence characteristics should play an important role in the observed differences in the biochemical characteristics of the XFPK and XPK homologs. The identified conserved indels are present in different regions of the XFPK protein sequence, and their structural analysis reveals that all of the identified CSIs, except possibly two, are located in the surface loops of XFPK. Earlier work on conserved indels provides evidence that the genetic changes represented by such indels are essential for the proper functioning of the proteins in the CSI-containing organisms, and the removal of such CSIs has detrimental effect on the proper functioning of the concerned proteins [60]. The localization of CSIs within surface loops of the proteins has also been noted in a number of previous studies [30,61,62]. The surface loops in protein sequences constitute highly accessible regions of the protein and they are known to play important roles in mediating protein-protein and protein-ligand interactions [63]. In a number of cases, surface loops in protein sequences due to either enabling (i.e. facilitate interaction) or disabling (prevent interactions)

characteristics have been shown to play important role in determining the oligomeric state of proteins [63,64].

Much of the work on PKs thus far has focused on bifidobacteria. The functional enzyme (XFPK) in bifidobacteria is a dimer with its active site located in between the two subunits. Our analyses of the conserved indels found in the bifidobacterial PKs show that at least two of these CSIs (viz. CSI #7 (D461-E462) and CSI #1 (F567-N569)) are located at the subunit interface, and they are indicated to play a role in the formation/stabilization of the protein dimer by means of hydrogen bonding, salt bridge formation or by providing an additional surface for subunit interaction. Studies on the dimerization potentials of the monomeric XFPK proteins, which either contained or lacked these CSIs, show that the docking scores for the XFPK monomers, which contained the CSIs were consistently higher in comparison to those obtained with the corresponding proteins that lacked these two CSIs. These results support the hypothesis that at least some of the CSIs, which distinguish the XFPK homologs of bifidobacteria from the XPK homologs, play an important role in the formation/stabilization of the dimeric form of the XFPK enzyme. In contrast to bifidobacteria, very limited work has been carried out on PKs from other bacteria and no reliable information is available concerning the oligomeric state of the functional XPK enzyme. Because the latter proteins are lacking the CSIs involved in the formation/stabilization of the dimeric protein, it is possible that the PK enzymes found in other microbes may function as monomers, or that the dimers formed in these cases are less stable, thus affecting the ability of the enzyme to bind to different substrates (viz. X5K and F6P). However, besides the CSIs that are indicated to be involved in dimer formation/stabilization, a number of other CSIs differentiating XFPK and XPK homologs are present in other parts/locations in the protein, and their influence on the overall functioning of the XFPK, including its regulation or its ability to recognize both X5K and F6P, remains to be explored. Further work on understanding the functional significance of different identified CSIs on the biochemical activities of the XFPK/XPK homologs should prove very relevant and informative in these regards.

Lastly, our observation that the PKs from bifidobacteria are closely related to those found in the *Coriobacteriales*, and that the PK gene in bifidobacteria was likely acquired from the latter group of microbes by means of HGT shifts our focus to the order *Coriobacteriales*. It is of interest in this regard that similar to the bifidobacteria, members of the order *Coriobacteriales* are also commensal organisms and they constitute significant constituents of the gut microbiota in humans and other animals [32,65,66]. Further similar to the *Bifidobacteriales*, some members of the order *Coriobacteriales* viz. *Atopobium* and *Olsenella* are associated with periodontal/endodontic infections, and the species *Atopobium vaginae* is commonly found (~ 80% of the cases) in bacterial vaginosis [32,67,68]. The order *Coriobacteriales* is a part of the class *Coriobacteriia* [31,32,52]. However, of the two orders that are present in this class, only members from the order *Coriobacteriales* exhibit saccharolytic ability and are able to metabolize glucose and wide variety of other carbohydrates, producing lactate and acetic acid as the main metabolites [31,32]. In contrast, the other order, *Eggerthellales*, is entirely made up of assacharolytic organisms [31] and no PK homolog could be detected in these bacteria. It should also be noted that Tween 80, which is a constituent of the growth medium for bifidobacteria, also exhibits a stimulatory effect on the growth of various *Coriobacteriales* species [3,8,32]. Thus, members of the order *Bifidobacteriales* and *Coriobacteriales* are very similar to each other in terms of their ecological niches, pathogenicity profiles, as well as their ability to utilize different carbohydrates and the metabolite end products produced [3,5,6,8,32,66,69]. In view of these observations and the remarkable similarity in the sequences of PK homologs from these two orders of bacteria, including the shared presence of large numbers of highly specific conserved indels, it is quite likely that the PK homologs from *Coriobacteriales*, similarly to the bifidobacteria, may

also be able to recognize and metabolize both X5K and F6P as substrates. Thus, it is possible that *Coriobacteriales* may constitute another group of microbes which are able to metabolize carbohydrates via the "bifid shunt". Further biochemical investigations in this regard should be of much interest.

Supporting information

S1 Fig. Sequence alignment files of PKs showing other conserved indels that are uniquely shared characteristic of the bifidobacteria and/or *Coriobacteriales* homologs.
(PDF)

Acknowledgments

We thank Seema Alnazar for providing technical assistance in this work. This work was supported by a research grant from the Natural Science and Engineering Research Council of Canada.

Author Contributions

Conceptualization: RSG AN BK.

Data curation: RSG AN BK.

Formal analysis: RSG AN BK.

Funding acquisition: RSG.

Investigation: RSG AN BK.

Methodology: RSG AN BK.

Project administration: RSG.

Resources: RSG.

Software: RSG BK.

Supervision: RSG.

Validation: RSG AN BK.

Visualization: RSG AN BK.

Writing – original draft: RSG AN.

Writing – review & editing: RSG AN BK.

References

1. Ventura M, Turroni F, Lugli GA, van Sinderen D. Bifidobacteria and humans: our special friends, from ecological to genomics perspectives. J Sci Food Agric.2014; 94: 163–168. doi: 10.1002/jsfa.6356 PMID: 23963950
2. Turroni F, van Sinderen D, Ventura M. Genomics and ecological overview of the genus *Bifidobacterium*. Int J Food Microbiol.2011; 149: 37–44. doi: 10.1016/j.ijfoodmicro.2010.12.010 PMID: 21276626
3. Biavati B: Family I. Bifidobacteriaceae Stackebrandt, Rainey and Ward-Rainey 1997, 487^{VP}. In *Bergey's Manual of Systematic Bacteriology, Volume 5, The Actinobacteria*. Edited by Whitman W, Goodfellow M, Kampfer P, Busse HJ, Trujillo ME, Ludwig W et al. New York: Springer; 2012:171.
4. Biavati B, Vescovo M, Torriani S, Bottazzi V. Bifidobacteria: histroy, ecology, physiology and applications. Ann Microbiolo.2000; 50: 117–131.

5. Pokusaeva K, Fitzgerald GF, van Sinderen D. Carbohydrate metabolism in Bifidobacteria. Genes Nutr.2011; 6: 285–306. doi: 10.1007/s12263-010-0206-6 PMID: 21484167
6. Palframan RJ, Gibson GR, Rastall RA. Carbohydrate preferences of Bifidobacterium species isolated from the human gut. Curr Issues Intest Microbiol.2003; 4: 71–75. PMID: 14503691
7. de Vries W, Gerbrandy SJ, Stouthamer AH. Carbohydrate metabolism in Bifidobacterium bifidum. Biochim Biophys Acta.1967; 136: 415–425. PMID: 6048259
8. Biavati B, Mattarelli P: Genus I. Bifidobacterium Orla-Jensen 1924, 472^{AL}. In *Bergey's Manual of Systematic Bacteriology*, *Volume 5*, *The Actinobacteria*. Edited by Whitman W, Goodfellow M, Kampfer P, Busse HJ, Trujillo ME, Ludwig W et al. New York: Springer; 2012:171–206.
9. Bottacini F, Milani C, Turroni F, Sanchez B, Foroni E, Duranti S, et al. Bifidobacterium asteroides PRL2011 Genome Analysis Reveals Clues for Colonization of the Insect Gut. PLoS One.2012; 7.
10. Cronin M, Ventura M, Fitzgerald GF, van Sinderen D. Progress in genomics, metabolism and biotechnology of bifidobacteria. Int J Food Microbiol.2011; 149: 4–18. doi: 10.1016/j.ijfoodmicro.2011.01.019 PMID: 21320731
11. Milani C, Turroni F, Duranti S, Lugli GA, Mancabelli L, Ferrario C, et al. Genomics of the Genus Bifidobacterium Reveals Species-Specific Adaptation to the Glycan-Rich Gut Environment. Appl Environ Microbiol.2015; 82: 980–991. doi: 10.1128/AEM.03500-15 PMID: 26590291
12. Turroni F, Bottacini F, Foroni E, Mulder I, Kim JH, Zomer A, et al. Genome analysis of *Bifidobacterium bifidum* PRL2010 reveals metabolic pathways for host-derived glycan foraging. Proc Natl Acad Sci U S A.2010; 107: 19514–19519. doi: 10.1073/pnas.1011100107 PMID: 20974960
13. Sanchez B, Delgado S, Blanco-Miguez A, Lourenco A, Gueimonde M, Margolles A. Probiotics, gut microbiota, and their influence on host health and disease. Mol Nutr Food Res.2016; 00: 1–15.
14. Scardovi V, Trovatelli LD. The fructose-6—phosphate shunt as peculiar pattern of hexose degradation in the genus *Bifidobacterium*. Ann Microbiol Enzimol.1965; 15: 19–29.
15. Grill JP, Crociani J, Ballongue J. Characterization of fructose 6 phosphate phosphoketolases purified from Bifidobacterium species. Curr Microbiol.1995; 31: 49–54. PMID: 7767228
16. Takahashi K, Tagami U, Shimba N, Kashiwagi T, Ishikawa K, Suzuki E. Crystal structure of Bifidobacterium Longum phosphoketolase; key enzyme for glucose metabolism in Bifidobacterium. FEBS Lett.2010; 584: 3855–3861. doi: 10.1016/j.febslet.2010.07.043 PMID: 20674574
17. Sanchez B, Zuniga M, Gonzalez-Candelas F, de los Reyes-Gavilan CG, Margolles A. Bacterial and eukaryotic phosphoketolases: phylogeny, distribution and evolution. J Mol Microbiol Biotechnol.2010; 18: 37–51. doi: 10.1159/000274310 PMID: 20068356
18. Meile L, Rohr LM, Geissmann TA, Herensperger M, Teuber M. Characterization of the D-xylulose 5-phosphate/D-fructose 6-phosphate phosphoketolase gene (xfp) from Bifidobacterium lactis. J Bacteriol.2001; 183: 2929–2936. doi: 10.1128/JB.183.9.2929-2936.2001 PMID: 11292814
19. Orban JI, Patterson JA. Modification of the phosphoketolase assay for rapid identification of bifidobacteria. J Microbiol Methods.2000; 40: 221–224. PMID: 10802138
20. Suzuki R, Katayama T, Kim BJ, Wakagi T, Shoun H, Ashida H, et al. Crystal structures of phosphoketolase: thiamine diphosphate-dependent dehydration mechanism. J Biol Chem.2010; 285: 34279–34287. doi: 10.1074/jbc.M110.156281 PMID: 20739284
21. Glenn K, Smith KS. Allosteric regulation of Lactobacillus plantarum xylulose 5-phosphate/fructose 6-phosphate phosphoketolase (Xfp). J Bacteriol.2015; 197: 1157–1163. doi: 10.1128/JB.02380-14 PMID: 25605308
22. Xiong W, Lee TC, Rommelfanger S, Gjersing E, Cano M, Maness PC, et al. Phosphoketolase pathway contributes to carbon metabolism in cyanobacteria. Nat Plants.2015; 2: 15187. doi: 10.1038/nplants.2015.187 PMID: 27250745
23. Yevenes A, Frey PA. Cloning, expression, purification, cofactor requirements, and steady state kinetics of phosphoketolase-2 from Lactobacillus plantarum. Bioorg Chem.2008; 36: 121–127. doi: 10.1016/j.bioorg.2008.03.002 PMID: 18430452
24. Yin X, Chambers JR, Barlow K, Park AS, Wheatcroft R. The gene encoding xylulose-5-phosphate/fructose-6-phosphate phosphoketolase (xfp) is conserved among Bifidobacterium species within a more variable region of the genome and both are useful for strain identification. FEMS Microbiol Lett.2005; 246: 251–257. doi: 10.1016/j.femsle.2005.04.013 PMID: 15899413
25. Burge G, Saulou-Berion C, Moussa M, Allais F, Athes V, Spinnler HE. Relationships between the use of Embden Meyerhof pathway (EMP) or Phosphoketolase pathway (PKP) and lactate production capabilities of diverse Lactobacillus reuteri strains. J Microbiol.2015; 53: 702–710. doi: 10.1007/s12275-015-5056-x PMID: 26428921
26. Petrareanu G, Balasu MC, Vacaru AM, Munteanu CV, Ionescu AE, Matei I, et al. Phosphoketolases from Lactococcus lactis, Leuconostoc mesenteroides and Pseudomonas aeruginosa: dissimilar

sequences, similar substrates but distinct enzymatic characteristics. Appl Microbiol Biotechnol.2014; 98: 7855–7867. doi: 10.1007/s00253-014-5723-6 PMID: 24740691

27. Posthuma CC, Bader R, Engelmann R, Postma PW, Hengstenberg W, Pouwels PH. Expression of the xylulose 5-phosphate phosphoketolase gene, xpkA, from Lactobacillus pentosus MD363 is induced by sugars that are fermented via the phosphoketolase pathway and is repressed by glucose mediated by CcpA and the mannose phosphoenolpyruvate phosphotransferase system. Appl Environ Microbiol.2002; 68: 831–837. doi: 10.1128/AEM.68.2.831-837.2002 PMID: 11823225

28. Vlkova E, Nevoral J, Jencikova B, Kopecny J, Godefrooij J, Trojanova I, et al. Detection of infant faecal bifidobacteria by enzymatic methods. J Microbiol Methods.2005; 60: 365–373. doi: 10.1016/j.mimet.2004.10.012 PMID: 15649538

29. Biavati B, Mattarelli P: The family ***Bifidobacteriaceae***. In *The Prokaryotes*: *An Evolving Electronic Resource for the Microbiological Community*. Edited by Dworkin M, Falkow S, Rosenberg E, Schleifer KH, Stackebrandt E. New York: Springer-Verlag; 2006:322–382.

30. Zhang G, Gao B, Adeolu M, Khadka B, Gupta RS. Phylogenomic Analyses and Comparative Studies on Genomes of the Bifidobacteriales: Identification of Molecular Signatures Specific for the Order Bifidobacteriales and Its Different Subclades. Front Microbiol.2016; 7: 978. doi: 10.3389/fmicb.2016.00978 PMID: 27446019

31. Gupta RS, Chen WJ, Adeolu M, Chai Y. Molecular signatures for the class *Coriobacteriia* and its different clades; Proposal for division of the class *Coriobacteriia* into the emended order *Coriobacteriales*, containing the emended family *Coriobacteriaceae* and *Atopobiaceae* fam. nov., and *Eggerthellales* ord. nov., containing the family *Eggerthellaceae* fam. nov. Int J Syst Evol Microbiol.2013; 63: 3379–3397. doi: 10.1099/ijs.0.048371-0 PMID: 23524353

32. Clavel T, Lepage P, Charrier C: The Family *Coriobacteriaceae*. In *The Prokaryotes- Actinobacteria*. Edited by Rosenberg E, DeLong E, Lory S, Stackebrandt E, Thompson F. New York: Springer; 2014.

33. Gupta RS: Identification of Conserved Indels that are Useful for Classification and Evolutionary Studies. In *Bacterial Taxonomy*, *Methods in Microbiology Volume 41*. Edited by Goodfellow M, Sutcliffe IC, Chun J. London: Elsevier; 2014:153–182.

34. Jeanmougin F, Thompson JD, Gouy M, Higgins DG, Gibson TJ. Multiple sequence alignment with Clustal x. Trends Biochem Sci.1998; 23: 403–405. PMID: 9810230

35. Castresana J. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Mol Biol Evol.2000; 17: 540–552. PMID: 10742046

36. Kumar S, Nei M, Dudley J, Tamura K. MEGA: a biologist-centric software for evolutionary analysis of DNA and protein sequences. Brief Bioinform.2008; 9: 299–306. doi: 10.1093/bib/bbn017 PMID: 18417537

37. Jones DT, Taylor WR, Thornton JM. The rapid generation of mutation data matrices from protein sequences. Computer applications in the biosciences: CABIOS.1992; 8: 275–282. PMID: 1633570

38. Whelan S, Goldman N. A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. Mol Biol Evol.2001; 18: 691–699. PMID: 11319253

39. Buchan DW, Minneci F, Nugent TC, Bryson K, Jones DT. Scalable web services for the PSIPRED Protein Analysis Workbench. Nucleic Acids Res.2013; 41: W349–W357. doi: 10.1093/nar/gkt381 PMID: 23748958

40. Eswar N, Webb B, Marti-Renom MA, Madhushudan MS, Eramian D, Shen MY, Pieper U, Sali A.. Comparative protein structure modelling using MODELLER. Curr.Protoc.Bioinformatics. 2007.

41. Shen MY, Sali A. Statistical potential for assessment and prediction of protein structures. Protein Sci.2006; 15: 2507–2524. doi: 10.1110/ps.062416606 PMID: 17075131

42. Lovell SC, Davis IW, Arendall WB III, de Bakker PI, Word JM, Prisant MG, et al. Structure validation by Calpha geometry: phi,psi and Cbeta deviation. Proteins.2003; 50: 437–450. doi: 10.1002/prot.10286 PMID: 12557186

43. Colovos C, Yeates TO. Verification of protein structures: patterns of nonbonded atomic interactions. Protein Sci.1993; 2: 1511–1519. doi: 10.1002/pro.5560020916 PMID: 8401235

44. Luthy R, Bowie JU, Eisenberg D. Assessment of protein models with three-dimensional profiles. Nature.1992; 356: 83–85. doi: 10.1038/356083a0 PMID: 1538787

45. Bowie JU, Luthy R, Eisenberg D. A method to identify protein sequences that fold into a known three-dimensional structure. Science.1991; 253: 164–170. PMID: 1853201

46. Krissinel E, Henrick K. Inference of macromolecular assemblies from crystalline state. J Mol Biol.2007; 372: 774–797. doi: 10.1016/j.jmb.2007.05.022 PMID: 17681537

47. Pierce BG, Hourai Y, Weng Z. Accelerating protein docking in ZDOCK using an advanced 3D convolution library. PLoS ONE.2011; 6: e24657. doi: 10.1371/journal.pone.0024657 PMID: 21949741

48. Duhovny D, Nussinov R, Wolfson H: Efficient unbound docking of rigid molecules. In *Algorithms in Bioinformatics*. Edited by Guigo R, Gusfield D. Berlin: Springer; 2002:185–200.
49. Comeau SR, Gatchell DW, Vajda S, Camacho CJ. ClusPro: an automated docking and discrimination method for the prediction of protein complexes. Bioinformatics.2004; 20: 45–50. PMID: 14693807
50. Kozakov D, Brenke R, Comeau SR, Vajda S. PIPER: an FFT-based protein docking program with pairwise potentials. Proteins.2006; 65: 392–406. doi: 10.1002/prot.21117 PMID: 16933295
51. Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, et al. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic Acids Res.2013; 41: D590–D596. doi: 10.1093/nar/gks1219 PMID: 23193283
52. Zhi XY, Li WJ, Stackebrandt E. An update of the structure and 16S rRNA gene sequence-based definition of higher ranks of the class *Actinobacteria*, with the proposal of two new suborders and four new families and emended descriptions of the existing higher taxa. Int J Syst Evol Microbiol.2009; 59: 589–608. doi: 10.1099/ijs.0.65780-0 PMID: 19244447
53. Gao B, Gupta RS. Phylogenetic framework and molecular signatures for the main clades of the phylum Actinobacteria. Microbiol Mol Biol Rev.2012; 76: 66–112. doi: 10.1128/MMBR.05011-11 PMID: 22390973
54. Segata N, Bornigen D, Morgan XC, Huttenhower C. PhyloPhlAn is a new method for improved phylogenetic and taxonomic placement of microbes. Nat Commun.2013; 4: 2304. doi: 10.1038/ncomms3304 PMID: 23942190
55. den Besten G, van Eunen K, Groen AK, Venema K, Reijngoud DJ, Bakker BM. The role of short-chain fatty acids in the interplay between diet, gut microbiota, and host energy metabolism. J Lipid Res.2013; 54: 2325–2340. doi: 10.1194/jlr.R036012 PMID: 23821742
56. Maslowski KM, Vieira AT, Ng A, Kranich J, Sierro F, Yu D, et al. Regulation of inflammatory responses by gut microbiota and chemoattractant receptor GPR43. Nature.2009; 461: 1282–1286. doi: 10.1038/nature08530 PMID: 19865172
57. Fleige C, Kroll J, Steinbuchel A. Establishment of an alternative phosphoketolase-dependent pathway for fructose catabolism in Ralstonia eutropha H16. Appl Microbiol Biotechnol.2011; 91: 769–776. doi: 10.1007/s00253-011-3284-5 PMID: 21519932
58. Glenn K, Ingram-Smith C, Smith KS. Biochemical and kinetic characterization of xylulose 5-phosphate/fructose 6-phosphate phosphoketolase 2 (Xfp2) from Cryptococcus neoformans. Eukaryot Cell.2014; 13: 657–663. doi: 10.1128/EC.00055-14 PMID: 24659577
59. Ruiz L, Hidalgo C, Blanco-Miguez A, Lourenco A, Sanchez B, Margolles A. Tackling probiotic and gut microbiota functionality through proteomics. J Proteomics.2016; 147: 28–39. doi: 10.1016/j.jprot.2016.03.023 PMID: 27003613
60. Singh B, Gupta RS. Conserved inserts in the Hsp60 (GroEL) and Hsp70 (DnaK) proteins are essential for cellular growth. Mol Genet Genomics.2009; 281: 361–373. doi: 10.1007/s00438-008-0417-3 PMID: 19127371
61. Cherkasov A, Lee SJ, Nandan D, Reiner NE. Large-scale survey for potentially targetable indels in bacterial and protozoan proteins. Proteins.2006; 62: 371–380. doi: 10.1002/prot.20631 PMID: 16315289
62. Gupta RS. Impact of genomics on the understanding of microbial evolution and classification: the importance of Darwin's views on classification. FEMS Microbiol Rev.2016; 40: 520–553. doi: 10.1093/femsre/fuw011 PMID: 27279642
63. Akiva E, Itzhaki Z, Margalit H. Built-in loops allow versatility in domain-domain interactions: lessons from self-interacting domains. Proc Natl Acad Sci U S A.2008; 105: 13292–13297. doi: 10.1073/pnas.0801207105 PMID: 18757736
64. Itzhaki Z, Akiva E, Altuvia Y, Margalit H. Evolutionary conservation of domain-domain interactions. Genome Biol.2006; 7: R125. doi: 10.1186/gb-2006-7-12-r125 PMID: 17184549
65. Nicholson JK, Holmes E, Kinross J, Burcelin R, Gibson G, Jia W, et al. Host-gut microbiota metabolic interactions. Science.2012; 336: 1262–1267. doi: 10.1126/science.1223813 PMID: 22674330
66. Graf D, Di Cagno R, Fak F, Flint HJ, Nyman M, Saarela M, et al. Contribution of diet to the composition of the human gut microbiota. Microb Ecol Health Dis.2015; 26: 26164. doi: 10.3402/mehd.v26.26164 PMID: 25656825
67. Kinross JM, Darzi AW, Nicholson JK. Gut microbiome-host interactions in health and disease. Genome Med.2011; 3: 14. doi: 10.1186/gm228 PMID: 21392406
68. Thorasin T, Hoyles L, McCartney AL. Dynamics and diversity of the 'Atopobium cluster' in the human faecal microbiota, and phenotypic characterization of 'Atopobium cluster' isolates. Microbiology.2015; 161: 565–579. doi: 10.1099/mic.0.000016 PMID: 25533445
69. Serino M, Luche E, Gres S, Baylac A, Berge M, Cenac C, et al. Metabolic adaptation to a high-fat diet is associated with a change in the gut microbiota. Gut.2012; 61: 543–553. doi: 10.1136/gutjnl-2011-301012 PMID: 22110050

CHAPTER 5

Novel Sequence Feature of SecA Translocase Protein Unique to the Thermophilic Bacteria: Bioinformatics Analyses to investigate their Potential Roles

PREFACE

This chapter highlights the identification of two large CSIs in SecA proteins that are distinctive molecular characteristics of the members of thermophilic and hyperthermophilic group of bacteria. This chapter also describes the detailed phylogenetic and structural studies using bioinformatic approaches to examine the evolution and structural features of these large CSIs. MD studies show that residues from one of the CSI play a role in mediating a conserved network of water molecules in *Thermotoga maritima* SecA at high temperature. My contributions towards the completion of this chapter includes the construction of phylogenetic tree based on SecA protein sequences, analysis of the amino acid compositions, homology modelling of SecA proteins without CSIs and mapping of the structural features of the identified CSIs using the available structure and homology models, and molecular dynamics (MD) simulations studies at various temperature settings and analysis of the MD trajectories. In addition, I was involved in the writing of the drafts and revision of the manuscript, and in the construction of the main and supplemental figures provided.

microorganisms

MDPI

Article

Novel Sequence Feature of SecA Translocase Protein Unique to the Thermophilic Bacteria: Bioinformatics Analyses to investigate their Potential Roles

Bijendra Khadka[1], Dhillon Preusad [1] and Radhey S. Gupta [1,*]

[1] Department of Biochemistry and Biomedical Sciences, McMaster University, Hamilton, ON L8N 3Z5, Canada; khadkab@mcmaster.ca, persaudk@mcmaster.ca

* Correspondence: gupta@mcmaster.ca; Tel.: +(905)-525-9140 (ext. 22639)

Received: date; Accepted: date; Published: date

Abstract: SecA is an evolutionarily conserved protein that plays an indispensable role in the secretion of proteins across bacterial cell membrane. Comparative analyses of SecA homologs have identified two large conserved signature inserts (CSIs) that are unique characteristics of thermophilic bacteria. A 50 aa conserved insert in SecA is exclusively present in the SecA homologs from the orders *Thermotogales* and *Aquificales*, while a 76 aa insert in SecA is specific for the *Thermales* species and *Hydrogenibacillus schlegelii*. Phylogenetic analyses on SecA sequences show that the shared presence of these CSIs in unrelated groups of thermophiles is not due to lateral gene transfers, but instead these large CSIs have likely originated independently in these lineages due to their advantageous function. Both these CSIs are located in SecA protein in surface exposed region within the ATPase domain. To gain insights into the functional significance of the 50 aa CSI in SecA, molecular dynamics (MD) simulations were performed at two different temperatures using ADP bound *Thermotogae maritima* SecA. These analyses have identified a conserved network of water molecules near the 50 aa insert in which the GLU185 residue from the CSI is found to play a key role towards stabilizing these interactions. The results provide evidence for the possible role of the 50 aa CSI in stabilizing the binding interaction of ADP/ATP, which is required for SecA function. Additionally, the surface-exposed CSIs in SecA, due to their potential to make novel protein-protein interactions, could also contribute to the thermostability of SecA from thermophilic bacteria.

Keywords: novel sequence features in SecA from thermophilic bacteria; phylogenetic analysis; conserved signature indels; molecular dynamics simulations of *Thermotoga maritima* SecA; conserved water molecules

1. Introduction

Thermophilic organisms (bacteria) are of great scientific interest due to their ability to grow at temperatures well above 60°C [1, 2]. The thermostability of the proteins from model organisms such as *Thermotoga maritima,* which can survive within a wide temperature range of 55-90°C, has been an area of intense research interest [3-9]. Thermostability of the protein has practical applications in industrial settings, biotechnologies, and bio-refining [8, 10-15]. Specifically, within industrial settings, the higher temperature stability of these protein catalysts allows for reactions at higher temperatures resulting in decreased contamination concerns and overall faster reaction speeds [16]. An example in this regard includes widespread use of enzyme *Thermus aquaticus* (*Taq*) polymerase in the technique polymerase chain reaction (PCR) [17]. Several comparative studies have shown that the thermostability of the proteins from thermophilic groups of organisms can be attributed to various characteristics [4-7, 9, 18, 19]. One prevalent characteristic is the increase in the presence of

ion-pair interactions in thermophilic organisms [18, 20]. The increase in ion-pair interactions is due to a higher composition of charged amino acids such as lysine (LYS), arginine (ARG), glutamic acid (GLU) and asparagine (ASN) in the proteins from thermophilic bacteria when compared to those from mesophilic bacteria [18, 20]. Sequence and structural characteristics such as presence of insertions and deletions, proline substitutions, closer packing of water-accessible surface residues, and increase in helical contents and hydrogen bonds has also been suggested to contribute towards increase in the thermostability of thermophilic proteins [21-25]. Evidently, the characteristics providing thermostability to proteins can exist in various forms and further understanding of protein features and characteristics that likely contribute towards the thermostability of proteins is of much interest [20, 26, 27].

Within the domain Bacteria, hyperthermophilic organisms are mainly present in three bacterial phyla viz. Aquificae, Deinococcus-Thermus, and Thermotogae [16, 28-32]. The members from these phyla, which contain some of the most hyperthermophilic organisms known (e.g. *Thermotogae maritima, Thermus aquaticus,* and *Aquifex aeolicus*), are notable for their thermostable enzymes [8, 10, 11, 16, 33]. However, of these three phyla, while the phylum Aquificae is primarily comprised of thermophilic-hyperthermophilic organisms [28, 30], in the other two phyla, hyperthermophilicity is a shared characteristics of species from only some orders (viz. *Thermotogales* and *Thermales*) of these bacteria [32, 34-37]. Our earlier comparative analyses on protein sequences from members of these three phyla have identified large numbers of molecular markers in the form of conserved signature indels (CSIs) in different proteins which are specifically shared characteristics of different members from each of these three phyla (viz. Thermotogae, Aquificae and Deinococcus-Thermus) of bacteria and their suborders [13, 30, 34, 38-42]. Although the CSIs that we have previously identified provide very useful means for distinguishing members of these phyla from each other, as well as other groups of bacteria, it is unclear if any of these genetic/biochemical changes played any role in the thermostability of these organisms. However, in the present study we describe the identification and analysis of two large CSIs in the homologs of SecA proteins, which due to their unique shared presence in members from two different phyla of hyperthermophilic bacteria are indicated to play important role in the thermostability of this protein.

SecA is a conserved ATPase whose homologs are well-preserved among all bacteria [43, 44]. It constitutes a major molecular motor for the ATP driven secretion of pre-proteins from the bacterial transmembrane protein translocation complex SecYEG [45, 46]. Previous biochemical and structural studies on SecA from thermophilic bacteria have contributed significantly towards understanding the overall architecture and function of the protein [47-49]. However, it remains unclear whether the presence of any unique sequence feature(s) in SecA contributes to its stability or for the functioning of SecA translocase at high temperatures. In the present study, we describe the identification and analysis of two large CSIs in the SecA proteins which are uniquely found in the SecA homologs from two different phyla of hyperthermophilic bacteria. One of these CSIs, consisting of an insert of 50 aa in a conserved region is specifically found in the SecA homologs from all members of the order *Thermotogales* (phylum Thermotogae) and *Aquificales* (phylum Aquificiae), while another large 76 aa insert is a uniquely shared characteristics of the SecA homologs from the order *Thermales* (phylum Deinococcus-Thermus) and in *Hydrogenibacillus schlegelii*, a thermophilic bacterium from the phylum Firmicutes. The shared presence of these large and unique sequence features in the SecA proteins from only the hyperthermophilic members of these phyla strongly suggests that the identified genetic/biochemical changes are related to the thermophilic characteristics of these organisms and they should be playing important role in the thermostability of this protein.

We report here the results of phylogenetic studies examining the evolution of these large CSIs in the thermophilic organisms, as well as sequence compositions, structural features and location of these CSIs in the SecA protein structure using the homology models and available structural information. Lastly, to gain some insights into the functional significance of the 50 aa CSI in SecA protein (which is uniquely found in members of the orders *Thermotogae* and *Aquificales*), we have performed molecular dynamics (MD) simulation at two different temperatures (303.15K and 363.15K) using the available SecA structure (*Tm*SecA) from *T. maritima,* which contains this large

CSI. The results from MD simulation studies identify a conserved network of water molecules whose interaction with the bound nucleotide in the SecA protein is mediated/stabilized by certain conserved residues from this CSI in *Tm*SecA. The significance of these observations in the context of the specificity of the identified CSIs for thermophilic organisms are discussed.

2. Materials and Methods

2.1. Identification of Conserved Signature Indels (Insertions/Deletions) and Phylogenetic Analysis

The described CSIs in the SecA proteins were identified as described in earlier work [40, 50, 51]. In brief, BLASTp searches were carried out on the SecA protein sequences from Thermotogae and other thermophilic organisms. Based on these BLASTp searches, sequences of SecA proteins were retrieved from 10-15 organisms representing different thermophilic phyla as well as similar number of sequences from members of other bacterial phyla. Multiple sequence alignments (MSAs) of the retrieved sequences were created using ClustalX 2.1 [52, 53]. These sequence alignments were examined for the presence of conserved inserts or deletions (i.e. indels), which were specifically found in thermophilic organisms, and which were flanked on both sides by at least 5 conserved residues in the neighboring 30-40 aa. More detailed BLASTp searches on the sequence regions containing the indels of interest and their flanking 40-50 aa were then conducted against the NCBI non-redundant (nr) database to determine the specificity of the identified indels. For the indels of interest, the signature files shown here were created using SIG_CREATE and SIG_STYLE program (from www.GLEANS.net). Unless otherwise indicated, all of the reported CSIs are specific for the group of interest and similar CSIs were not observed in homologs from any other bacterial species within the top 500 BLASTp hits examined. A phylogenetic tree was produced based on SecA sequences from multiple groups of bacteria. These sequences were obtained from a BLASTp search of the SecA sequence for the different taxonomic groups of interest. Based on the sequence alignment of SecA protein from different species, a maximum-likelihood (ML) tree phylogenetic tree based on 100 bootstrap replications was constructed using MEGA6 program [54] employing the Whelan and Goldman model (WAG) of protein sequence evolution [55].

2.2. Homology Modelling of SecA homologs and Structural Analysis of CSIs

The homology models of the *Thermotoga maritima* and *Thermus thermophilus* SecA proteins lacking the 50 aa CSI were created using an in-house pipeline, "GlabModeller", as described in earlier work [56-60]. The available crystal structures of SecA from *T. maritima* (PDB ID: 4YS0) and *T. thermophilus* (PDB ID: 2IPC) were used as templates [47, 48]. In brief, for homology modeling, the sequence alignments between target and template proteins were carried out using the align2D module from the MODELLER, which is integrated and streamlined in GlabModeller tool. The resulting alignments were carefully analyzed and modified manually to ensure the reliability of the location of CSIs. For each target protein, 500 models were ranked on the basis of their Discrete Optimized Protein Energy (DOPE) scores [61]. Selected models were then refined using ModRefiner [62]. The stereo-chemical properties of the final models were assessed using three independent servers which include RAMPAGE [63], ERRAT [64], PROSA [65, 66] and VERIFY3D [67]. These applications utilize a dataset of refined structures to evaluate the statistical significance of the model conformation, location, environment of each amino acid sequence and overall structural stability. The resultant models were then used to explore the structural changes associated with the CSI. The superimposition of the validated models with the template structures was carried out using PyMOL (Version 1.7.4; Schrödinger, LLC.) to examine the structure and location of identified CSIs in the SecA structures.

2.3. Molecular Dynamics Simulations

The all-atom molecular dynamics simulations were performed using (Groningen machine for chemical simulations) GROMACS 5.1.2 software [68, 69] with the all-atom CHARMM36 force field for SecA, ADP, ions together with 3-points (TIP3P) water model and added ionic strength to mimic

the physiological environment [70]. The atomic coordinates of the *T. maritima* SecA (PDB ID: 4YS0) resolved at 1.85 Å were obtained from protein data bank (PDB) and the homology model of *Tm*SecA lacking CSI was utilized. The potential energy of the system was minimized using a 50,000-step steepest descendent method to relax the system and to avoid any steric clashes. Missing amino acid residues in the crystal structure of *Tm*SecA were identified and fixed using the MODELLER implemented in the GlabModeller tool. After energy minimization, the system was equilibrated with isothermal-isochoric/NVT (constant number of particles, volume, and temperature) ensemble and then 100 ns of MD simulation in the isothermal-isobaric/NPT (constant number of particles, pressure, and temperature) ensemble using the Nose-Hoover thermostat and Parrinello-Rahman barostat [71-73]. Comparative analysis of the difference in binding affinity of ADP towards *T. maritima* SecA with CSI and without CSI was carried out. The *T. maritima* SecA with ADP was simulated under two different reference temperatures of 303.15K and 363.15K. These large temperature gaps were selected to investigate the effect of temperature on the *Tm*SecA dynamics during the simulation period. The root mean square deviation (RMSD) and hydrogen bond interaction calculations were carried out using the GROMACS utilities. All MD simulation runs were carried out using our local GROMACS certified graphical processing unit (GPU) accelerated high-performance computing system obtained from EXXACT Corporation [74]. A total of 1000 snapshots were extracted for every 100 ps from the 100 ns MD trajectories to analyze the dynamics of water molecules near CSI containing region. The analyses of the structures obtained from trajectories were carried out using the various utilities of the GROMACS [68], VMD [75] and PyMOL (www.pymol.com).

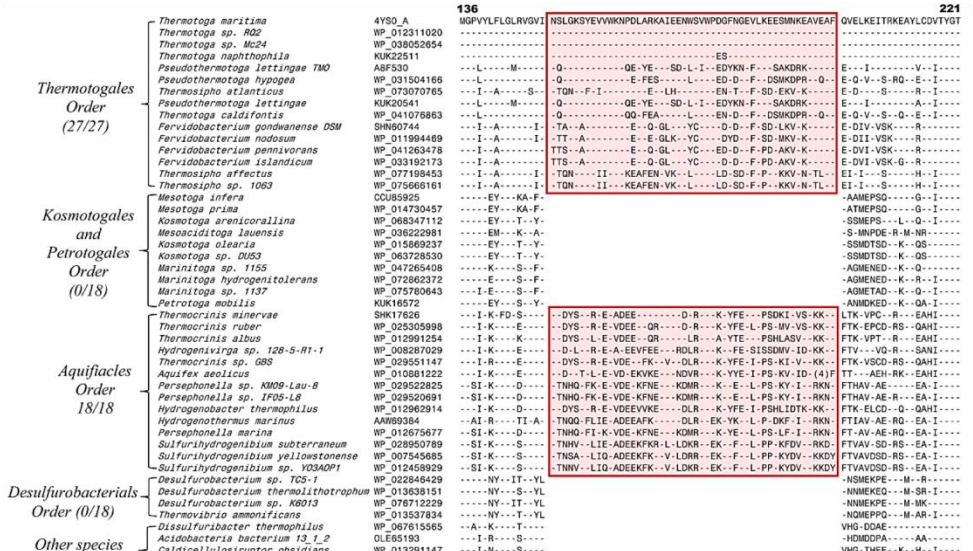
3. Results

3.1. Identification of Conserved signature Indels in SecA homologs from Thermotogales, Aquificales, and Thermales and their Phylogenetic Implications

Our comparative analysis of SecA protein sequences from different bacterial groups has identified several CSIs in SecA protein that are specific for particular groups/taxa of bacteria. However, the present study focuses on our identification of two large CSIs in the SecA protein, which are uniquely found in different homologs from the thermophilic-hyperthermophilic phyla of bacteria. The first of these CSIs shown in Figure 1 is a 50 aa insertion in SecA homologs that is uniquely shared by members of the order *Thermotogales* and *Aquificales*. As seen from Figure 1, within the phylum Thermotogae, this CSI is a shared characteristics of all members from the order *Thermotogales*, but it is not found in any of the species from the orders *Peterotogales* and *Kosmotogales*. It should be noted in this regard, that within the phylum Thermotogae, the order *Thermotogales* encompasses all of the thermophilic-hyperthermophilic organisms, whereas the other two orders lacking this CSI are comprised of mesophilic organisms [13, 34, 37, 76]. Thus, within the phylum Thermotogae, this CSI is uniquely found in the organisms which are thermophilic-hyperthermophilic [39, 40]. Interestingly, in addition to the *Thermotogales*, this large CSI is also commonly shared by different species belonging to the order *Aquificales* from the phylum Aquificae, which are also comprised exclusively of hyperthermophilic organisms. However, within this phylum members belonging to the order *Desulfurobacteriales*, which are strict anaerobes [30, 77] do not contain this large insert.

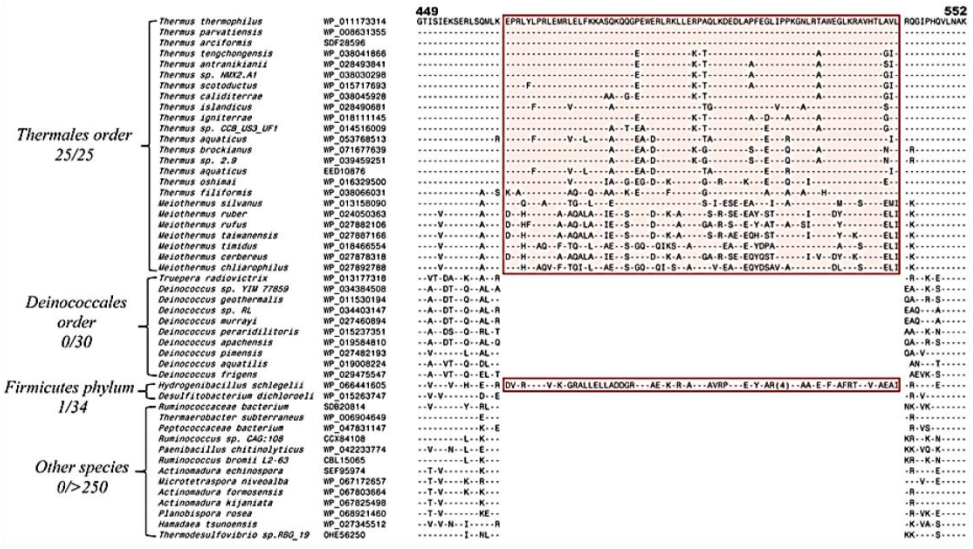
The second large CSI in SecA protein that we have identified is a 76 aa insertion in a conserved region, (see Figure 2), which is commonly shared by all SecA homologs from members of the order *Thermales* belonging to the phylum Deinococcus-Thermus. The phylum "Deinococcus-Thermus" contains two extensively studied orders of extremophilic microorganisms i.e. *Deinococcales* and *Thermales* [32, 42]. Of these two orders, the order *Thermales* is comprised exclusively of organisms that are thermophilic and hyperthermophilic [32, 42] whereas the member of the order *Deinococcales* are known for their high degree of radiation resistance [35, 42, 78] Interestingly, this large insert in SecA is found only in different members of the order *Thermales* but not in any of the homologs from *Deinococcales*. Further, in addition to the *Thermales* species, this insert in SecA is also commonly

194 shared by the species *Hydrogenibacillus schlegelii,* which is a thermophilic bacterium belonging to the
195 phylum Firmicutes [79-81]. Thus, both these large CSIs in SecA are only found in members of
196 different main orders/phyla of bacteria that contain thermophilic-hyperthermophilic organisms.
197 Except for the thermophilic-hyperthermophilic organisms, these inserts are not present in any other
198 SecA homologs. Thus, the indicated CSIs are further examined in the context of their role in
199 thermostability.



200

201 **Figure 1.** Excerpts from the sequence alignment of SecA protein showing a 50 aa conserved insert
202 that is a distinctive characteristic of species from the orders *Thermotogales* and *Aquificales* but absent
203 in the homologs from all other bacteria. The dashes (-) in the sequence alignment denote sequence
204 identity with the amino acid shown on the top line. The accession numbers of the protein sequences
205 are provided in the second column and numbers on top of the sequence alignment indicate the
206 position of this sequence in *T. maritima*.



207

Figure 2. Partial sequence alignment of SecA showing a 76 aa conserved insert that is a unique characteristic of species from the order *Thermales* and *H. schlegelii* but absent from the SecA homologs from all other bacteria. Except for *H. schlegelii*, no other Firmicutes species contained this insert. Other information concerning the sequence alignment is the same as in Figure 1.

3.2. Phylogenetic Analysis of the SecA proteins to Investigate the Shared Presence of CSIs

As both these large CSIs are present in only the thermophilic members from two different phyla of bacteria, it was of much interest to investigate how the shared presence of these genetic changes in two distinct groups/phyla of bacteria could be explained. Based on earlier work horizontal gene transfers are indicated to occur frequently between members of the bacterial phyla that contain thermophilic-hyperthermophilic organisms [30, 37, 40, 82]. Thus, we have examined whether the shared presence of these CSIs in these two cases is due to horizontal transfers of SecA gene between the two groups of organisms which contain either the 50 aa or the 76 aa inserts. To investigate this, we have constructed a maximum-likelihood phylogenetic tree based on SecA homologs from different relevant bacterial groups/phyla (Figure 3). In this tree, members from the phyla Thermotogae and Aquificae form distinct clades and the two orders of these bacteria viz. *Thermotogales* and *Aquificales*, which contain the 50 aa insert are separated from each other by other members of these phyla, which lack the 50 aa insert. Similarly, members of the order *Thermales* and the Firmicutes species *H. schlegelli*, both of which contained the 76 aa insert also did not cluster together in the phylogenetic tree. Instead, members of the order *Thermales* branched with other members from the phylum Deinococcus-Thermus, whereas *H. schlegelli* branched within a cluster of other species from the phylum Firmicutes. If the shared presence of the 50 aa CSI and 76 aa CSI in the two indicated groups of organisms was due to horizontal gene transfers, then it was expected that the SecA genes from these CSI-containing organisms would have clustered together in the tree. However, as the observed branching pattern of the CSIs-containing organisms is contrary to this expectation, it strongly suggests that the shared presence of the large CSIs in either the order *Thermotogales* and *Aquificales,* or in *Thermales* and *H. schlegelli*, is not due to horizontal gene transfers. Instead the results obtained suggest that the genetic changes leading to these large inserts have likely occurred independently (i.e. convergent evolution) in these lineages due to their presumed selective advantage.

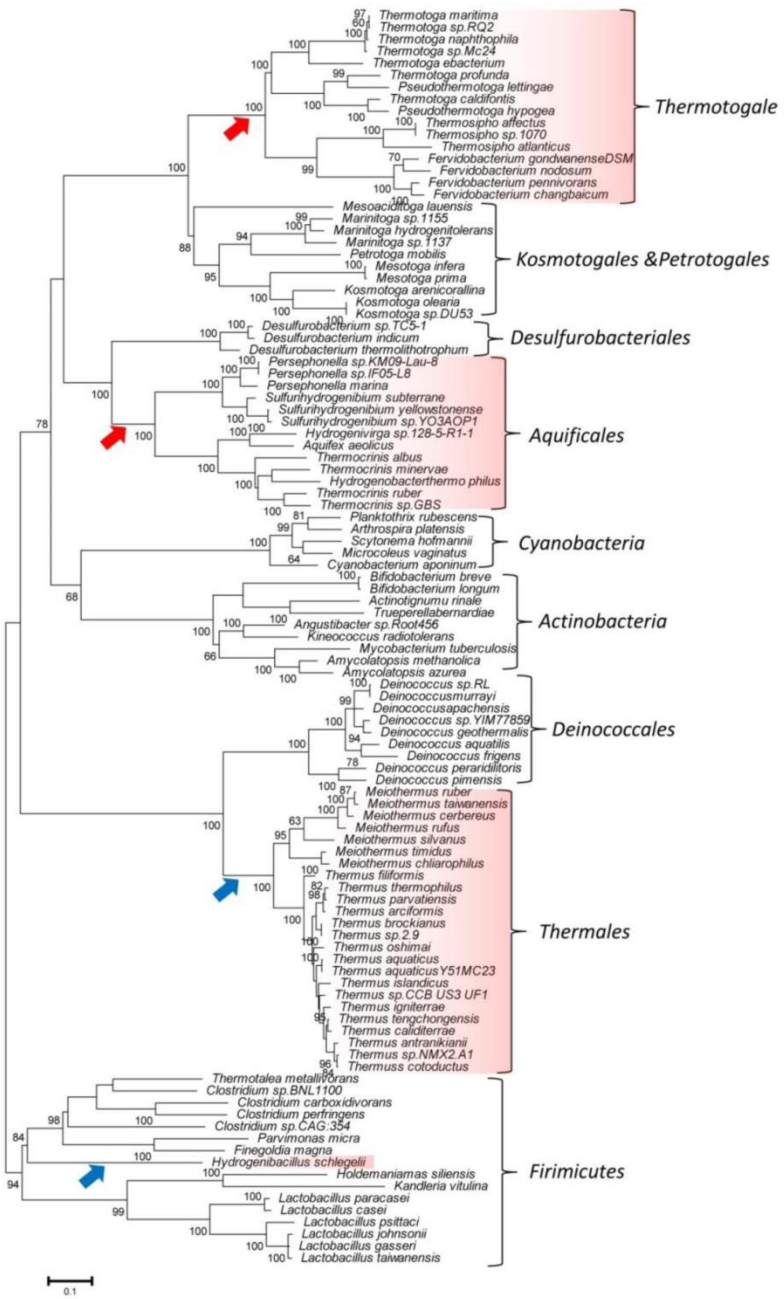


Figure 3. A maximum-likelihood phylogenetic tree based on SecA protein sequences from representative bacterial phyla. The groups of species containing the 50 aa CSI are marked by red arrows, whereas those containing the 76 aa CSI are denoted by blue arrows. The number on the nodes indicate bootstrap values for the observed groupings.

3.3. Computational Analysis of the CSIs in SecA proteins

As the two large CSIs in the SecA proteins are found exclusively in the bacterial groups/orders that consist entirely of thermophilic and hyperthermophilic organisms, it strongly suggests that they play some role in the thermostability of the SecA protein. Hence, exploration of the functional characteristics of these CSI could provide some insights into the thermostability of the SecA protein. Based on earlier studies, thermostability of proteins is enhanced by an increase in the charged amino acids that facilitate increased ion-pair interactions and stabilize the protein in high entropy environment [18, 20]. The results of our analyses (Figure S1) indicate that both the large CSIs in SecA protein contains a higher proportion of charged amino acids such as GLU, ARG, and LYS, which are known to facilitate increased ion-pair interactions.

The crystal structures of SecA are available from both mesophilic bacteria (*Bacillus subtilis* and *Mycobacterium tuberculosis*), and from thermophilic bacteria (*Thermus thermophilus* (PDB ID: 2IPC) and *Thermotogae maritima* (PDB ID: 4YS0)) [48, 49, 83-85]. The catalytic core of SecA protein is comprised of five functionally essential domains, which are shown in Figure 4A in the available crystal structure of the *T. maritima* SecA (*Tm*SecA) [47]. The different domains of SecA protein are highlighted using different color shades, cyan as a Nucleotide-binding domain (NBD1), magenta as Nucleotide-binding domain 2 (NBD2), red as a pre-protein binding domain (PPXD), yellow as Helical wind domain (HWD), and green as Helical scaffold domain (HSD). We have also created a homology model of the *Tm*SecA protein lacking the 50 aa CSI, using the protocol described in the Methods section, for the comparative analyses of structural features and location of this conserved insert (Figure 4B). As can be seen, the 50 aa conserved insert in *Tm*SecA specific for *Thermotogales* and *Aquificales* is located in a surface-exposed loop region of the NBD1, which forms a part of the ATP-binding site in the protein [47, 86]. The insert protrudes to form two additional β-strands at the periphery of the NBD1 domain with two short α-helices connected by a loop [49]. Although this surface-exposed CSI in the structure of *Tm*SecA is located in close proximity to the bound ADP molecule, it does not interact directly with bound ADP, nor it makes any contact with the SecY channel [49, 87]. We have also mapped the location of the 76 aa CSI specific for the order *Thermales* and *H. schlegelii* using the available crystal structure of SecA from *Thermus thermophilus* SecA (*Tt*SecA) (PDB: 2IPC) [48]. As can be seen from Figure 4C, this large CSI is also located in a surface-exposed loop region of the *Tt*SecA at the periphery of NBD2. However, unlike the 50 CSI in *Thermotogales* and *Aquificales*, this CSI is not in close proximity to the ADP-ATP binding site on the protein. However, due to its location, this CSI through its role in enabling intramolecular ionic interactions could be playing a role in stabilizing dimer formation at higher temperatures.

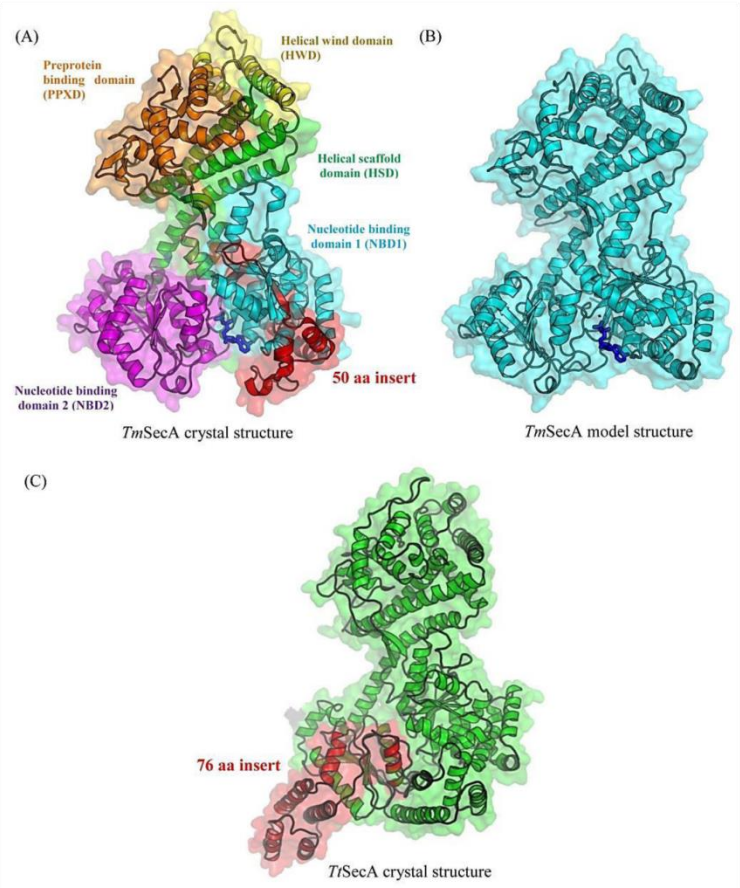


Figure 4. Cartoon and transparent surface representation of the 3D structure of A) *Thermotoga maritima* SecA (PDB ID: 4YS0) shown as green and the various structural domains present in this protein are colored and labeled, a B) homology model of *T. maritima* SecA lacking the 50 aa CSI. The bound ADP in the structure is shown as a blue stick. (C) Structure of SecA dimer from *Thermus thermophilus* (*Tt*SecA) (PDB ID: 2IPC). The 50 aa CSI and the 76 CSI are highlighted in red in these structures.

3.4. Molecular dynamics (MD) simulation studies of SecA containing 50 aa CSI specific for Thermotogales and Aquificales: analysis of TmSecA conformational stability and flexibility

In view of the location of the 50 aa CSI in the SecA of *Thermotogales* and *Aquificales* in close proximity to the ADP-ATP binding site, we have carried out MD simulation studies to gain some insights into the function of this CSI. In this regard, we have initially investigated the dynamic of the 50 aa CSI's flexibility from the trajectories obtained from the MD simulation studies on *Tm*SecA structure with CSI (+CSI) and without CSI (-CSI) at two different temperature settings. The detailed protocol of the system setup for MD simulations is described in the Methods section. In total four simulation runs were carried out using *Tm*SecA (+CSI) and *Tm*SecA (-CSI) each for 100 ns at two temperature settings of 303.15K (32 °C) and 363.15K (90°C). A preliminary analysis was carried out to analyze the overall deviation of *Tm*SecA (+CSI) and *Tm*SecA (-CSI) relative to their native structures as a function of time using the trajectories along the 100 ns time scale at these temperatures. At 303.15K, *Tm*SecA (+CSI) appear relatively more stable with an average RMSD of <

297 0.20 nm (2.0Å) (±0.04 nm) relative to 0.39 nm (3.9Å) (±0.08 nm) for *Tm*SecA (-CSI) (Figure S2A).
298 Similarly, at 363.15K, an average RMSD value of 0.34 nm (3.4 Å) (± 0.5 nm) for *Tm*SecA (+CSI)
299 relative to an average value of 0.36 nm (3.6Å) (±0.6 nm) for its insertion truncated homolog *Tm*SecA
300 (-CSI) (Figure S2B). Although at a high temperature the differences in RMSD values are minimal,
301 *Tm*SecA (-CSI) lacking the CSI showed a much higher degree of fluctuation, in comparison to the
302 protein with the CSI, indicating that the *Tm*SecA (+CSI) was more stable at the higher temperature
303 over the simulation period.

304 *3.5. Identification of Conserved CSI-mediated water network in TmSecA*

305 The crystal structure of *Tm*SecA with ADP bound contains a number of bound water molecules
306 near the ADP-binding site [47]. Of these water molecules, few forms an intermediate interaction
307 between the adenine group of ADP to the backbone of the residues (GLU 185 and VAL 186) from the
308 50 aa CSI in *Tm*SecA which is located near the ADP binding site. This observation is of much
309 interest, as several earlier studies indicate that water molecules make significant contribution
310 towards binding affinity of the ligand or mediating protein-ligand complexes by forming bridges
311 between the protein and the ligand [88-93]. To investigate this, using the protocol described in the
312 Methods section, computational analyses of the MD simulation trajectories were carried out to
313 determine whether the 50 aa CSI in *Tm*SecA might be interacting with ADP molecule by forming
314 intermediate interactions with water molecules.
315 Initially, we analyzed the hydration of the ADP binding site by calculating the presence of a
316 number of water molecules within the 9Å from ADP during the entire simulation of 100 ns at
317 303.15K and 363.15K (Figure S3). At 303.15K, the increase in the number of water molecules
318 increased in case of *Tm*SecA (+CSI) after 50 ns of the simulation, and overall it contained more water
319 molecules (with an average of total 129 molecules) when compared with *Tm*SecA (-CSI) that
320 contained an average total of 103 water molecules. However, at 363.15K, the number of water
321 molecules around the ADP binding site for *Tm*SecA (+CSI) decreased after 50 ns of simulation time
322 to an average total of 103 water molecules, whereas their numbers in *Tm*SecA (-CSI) slightly
323 increases to an average total number of 112 water molecules.
324 Further analyses of the MD trajectories identifies a network of water molecules that shows high
325 degree of conservancy and stable occupancy near the loop residues (amino acid range 183 to 188)
326 from the 50 aa CSI and these water molecules formed interactions with the adenine group of ADP
327 similar to that observed in the crystal structure of *Tm*SecA [47]. In a series of snapshots extracted at
328 different time intervals from 100 ns MD trajectories of *Tm*SecA (+CSI) at 303.15K (Figure 5A) and
329 363.15K (Figure 5B), we show the coordinates of water molecules, which constantly occupy the
330 location near the backbone of residues (amino acids 185-188) from this insert. For comparisons,
331 crystallographically observed positions of the water molecules in the *Tm*SecA crystal structure are
332 superimposed in this figure and they are shown as magenta spheres. Any hydrogen bonds formed
333 between simulation water molecules (red and white spheres), and adenine group of ADP or residues
334 from the insert are shown as yellow dash lines. As can be seen, at both temperatures, a network
335 involving two to three water molecules are maintained throughout the simulation period. Although
336 most of the water molecules are highly mobile spending only a fraction of time in that position,
337 interestingly the other water molecules that displace them occupy the same position and forms a
338 network of water-mediated interactions that are highly similar to those observed in the crystal
339 structure of *Tm*SecA protein [47].

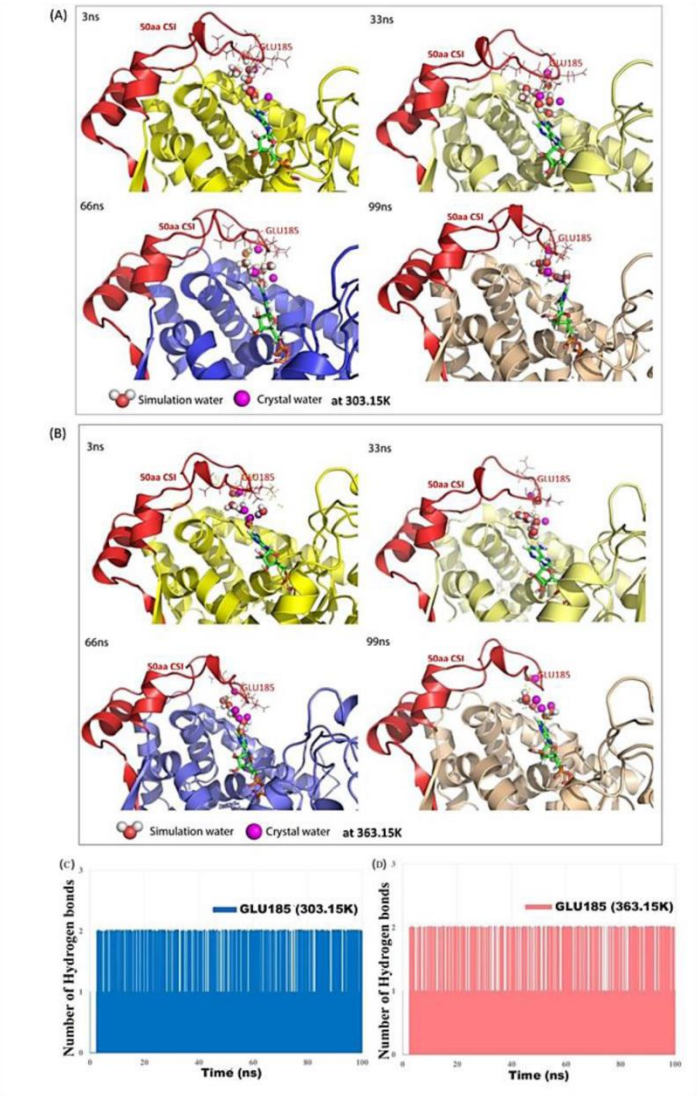


Figure 5. Snapshots from different time intervals extracted from the 100 ns MD trajectories of *Tm*SecA (+CSI) showing the coordinates of water molecules from the simulation (red and white spheres) that constantly occupy the location near the backbone of residue GLU185 from the 50 aa CSI in *Tm*SecA at **(A)** 303.15K, and **(B)** 363.15K. Each of the extracted structures representing the different time scale is represented with different colors and the times are indicated in the snapshots. The magenta spheres represent the coordinates for crystallographic water molecules which are from the corresponding region in the crystal structure of *Tm*SecA and they are superimposed to each of the extracted structures to show the occupancy of the simulation water molecules of that region. Any hydrogen bonds formed between simulation water molecules and adenine group of ADP or residues from the insert are shown as yellow dash lines. Other randomly picked snapshots at different time intervals that show the occupancy and interaction of water molecules with residue (GLU185) from 50 aa CSI and ADP in *Tm*SecA are provided in Figure S4A and S4B. **(C)** Trajectories showing the time evolution of the number of hydrogen bond interactions formed between GLU185 residue backbone atom (O) from the 50 aa CSI and water molecules that are within 4Å of GLU185 at 303.15K (blue) and **(D)** 363.15K (red), calculated over the 100 ns time period.

The time evolution of the hydrogen bond interactions calculated between GLU185 and water molecules that are within the 4Å of GLU185 over the course of 100 ns MD trajectory at 303.15K and 363.15K is shown in Figure 5C. It is of interest to note that the residue glutamic acid (GLU185) present in the loop region (residues 183-188) of this 50 aa insert is found to be conserved among SecA homologs from all members of the orders *Thermotogales* and *Aquificales*. As can be seen in Figure 5C, the backbone of GLU 185 residue is involved in bridging the two hydrogen bond interactions with water molecules throughout the simulation at both temperatures.

4. Discussion

Thermophilic bacteria often exhibit many evolutionary adaptations that aids to retain the function of their proteins at very high temperature [4-7, 9, 18]. These evolutionary adaptions such as ion-pair interactions, insertions and deletions, hydrogen bonds, and salt bridges ultimately result in increased intramolecular interactions and provide protein stability in the high entropy conditions [16, 18, 20, 94, 95]. In addition, a higher degree of close packing of water-accessible residues on the surface of the proteins, and contact orders has ben reported in thermophilic bacteria compared to their mesophilic homologs [9, 25]. However, despite the significant progress that has been made toward understanding of the structural peculiarities of thermophilic proteins, no unique or single mechanism has been found responsible, instead, a set of factors or their various combinations has been suggested to contribute towards the thermostability of a protein [8, 16, 94-98].

In the current work, we describe the identification of two large and unique molecular sequence features in the form of CSIs that are uniquely present in the SecA homologs from different thermophilic and hyperthermophilic members from all three main phyla of bacteria (viz. Thermotogae, Aquificae and Deinococcus-Thermus), which harbor most such organisms [16, 28-32]. Earlier studies on CSIs in different proteins show that these kind of rare genetic changes play important (or essential) roles in the functioning of the proteins within the CSI-containing organisms [78, 99] and any significant changes in these genetic characteristics are incompatible with their cellular function/growth [99]. In view of the specificities of the identified CSIs in the SecA homologs for thermophilic-hyperthermophilic organisms, and the importance of such genetic changes, it is of much interest to understand what unique functions/roles these large CSIs play in the indicated groups/phyla of thermophilic bacteria.

Results presented here show that the two large CSIs in SecA protein are commonly shared by members from two different groups/phyla of organisms. While the 50 aa CSI in SecA is a uniquely shared characteristics of the members of the order *Thermotogales* and *Aquificales*, the 76 aa CSI is specific for the order *Thermales* and *H. schlegelii*, a species from the phylum Firmicutes. In both cases, the identified CSIs are present in two unrelated groups of bacteria in the same location in SecA protein. Further, these CSIs are of the same (or similar) lengths and exhibit high degree of conservation in their amino acid sequences. As horizontal gene transfers among thermophilic organisms are indicated to occur frequently [30, 37, 40, 82], the simplest explanations to account for the presence of these large genetic characteristics in members from two unrelated phyla of bacteria would be that the genetic changes leading to either the 50 aa or 76 aa CSIs initially occurred in one of the lineages of thermophilic bacteria and then the SecA genes containing these CSIs were laterally transferred to the other phyla of thermophilic bacteria containing very similar CSIs. However, the phylogenetic analyses of SecA protein sequences do not support the view that the shared presence of these CSIs in the observed unrelated groups of thermophiles is due to lateral gene transfers. To account for the observed results, it is likely that the genetic changes leading to these large CSIs have occurred independently in two unrelated groups of thermophiles as a result of selective (pressure) advantageous functions of these insertions in the growth/survival of indicated groups of thermophilic organisms at high temperature. However, given the large sizes of both these CSIs and the observed high degree of sequence conservation within their sequences, the possibility that these genetic changes have occurred independently in different lineages appear surprising. Hence, another possibility to account for these results that can be considered is that instead of the horizontal transfer of the entire SecA genes, only the genetic exchanges or recombination of specific segments

of SecA genes containing these CSIs, have occurred between the indicated specific groups of thermophiles organisms [37]. As the rest of the SecA gene has evolved independently in these lineages, it will account for the distinct branching of organisms containing these CSIs in a phylogenetic tree based on SecA protein sequences.

It is of interest that within the phylum Aquificae, the 50 aa CSI is only shared by members of the order *Aquificales* but not by species from the order *Defulfurobacteriales*, which are also thermophilic. However, an important difference between members of the orders *Aquificales* and *Defulfurobacteriales* is that while the former order is comprised of species that are aerobic or microaerophilic and obtain energy from hydrogen or reduced sulphur compounds through molecular oxygen, the members of the latter order are strict anaerobes, that obtain energy by reduction of sulphate, nitrate, elemental sulphur, or other compounds by molecular hydrogen [30, 41, 100, 101]. Additionally, the members of these two orders are known to possess different metabolic pathways as their environments have different metabolic requirements [102]. These observations suggest that the 50 aa CSI in the SecA protein may confer selective advantage (thermostability) only under aerobic conditions.

However, irrespective of the evolutionary mechanisms that underlie the shared presence of these CSIs, based on the specific presence of these large CSIs in SecA homologs from only the thermophilic-hyperthermophilic organisms, it is strongly anticipated that these CSIs should be playing an important role in the thermostability of this essential protein at high temperature. Results from our structural analyses show that both these CSIs are located on the surface-exposed loops in a functionally important domain (NBD1) of the SecA protein. Of the two CSIs in SecA, the 50 aa CSI, which is specific for the order *Thermotogales* and *Aquificales*, is located in a surface-exposed loop that lies in close proximity to the ADP/ATP-binding site in the protein. Our comparison of the crystal structure of *Thermotoga maritima* SecA containing the 50 aa CSI with its homology model lacking the 50 aa CSI, has revealed that a number of water molecules forms an intermediate interaction between the residues from this CSI and ADP molecule. Although, the roles of conserved water molecules in mediating protein-ligand interactions have been increasingly recognized in recent years [89, 91-93, 103-106], our understanding of the significance of these conserved waters in SecA, or how other sequence features of the protein contribute towards their conservation, is limited. To investigate this aspect further, molecular dynamics (MD) simulations were carried out in this work to examine the structural dynamics of the 50 aa CSI in SecA protein and the water molecules found in its proximity in the crystal structure of *Tm*SecA at two different temperatures (303.15K, and 363.15K). The results from MD studies identify a network of highly stable water molecules that forms an intermediate interaction between the residues such as GLU185 from the 50 aa CSI and adenine group of ADP at both temperatures. Earlier studies have indicated that the hydrogen bonding capability of water molecules with charged amino acids contributes towards the stability of hydrogen bonds formed between coenzymes/cofactors like ADP/ATP and proteins [22, 107-111]. In view of these earlier studies, the high residence time of the water molecules inside the cavity formed by the 50 aa CSI and their forming a conserved hydrogen bonding network with some conserved residues from the CSI during the course of simulation, strongly suggests that the 50 aa CSI likely plays a role in maintaining the constant stable network of water molecules which likely plays a role in stabilization of the ADP-ATP molecules in the active site of the protein at high temperature. Since ATP binding and hydrolysis is essential for the functioning of SecA protein, it is possible that the observed conserved hydrogen bonding network between the water molecules, residues from the CSI, and ADP (ATP), helps in stabilizing the binding of ATP to the protein at high temperature.

However, the suggested role of the 50 aa CSI in stabilizing the binding of ADP/ATP to the protein at high temperature could be only one of the several factors by which this large CSI might be contributing towards the thermostability of the organisms for which they are specific. As noted earlier, these large CSI as well as the 76 aa CSI, specific to the member of the order *Thermales* and *H. schlegelli*, are present as surface-exposed loops in the structure of SecA protein. Based on earlier work, the surface loops in proteins are often involved in mediating novel protein-protein or protein-ligand interactions or in maintaining/stabilizing a specific oligomeric state of the proteins [56-58, 99, 112-116]. Earlier structural and functional studies also indicate that SecA can adopt a wide

variety of oligomeric states [117-120] and it is possible that the presence of these CSI could stabilize certain oligomeric forms that are of relevance for the functioning of this protein in thermophilic environment. In addition, most thermophilic and hyperthermophilic bacteria lacks SecB, a molecular chaperone protein which plays an important role in transferring pre-protein to SecA and SecYEG translocation system in most other bacteria [121-123]. Thus, it is likely that members of thermophilic and hyperthermophilic phyla utilize other chaperons or proteins for similar functions [122]. Although, other novel proteins that may be required for the functioning of SecA protein in thermophilic organisms have not yet been characterized, it is possible that the presence of the unique surface-exposed loops formed by the CSIs in the SecA proteins from thermophilic organisms could serve as a platform for the unique binding of these proteins with the SecA. Overall, the results provided in this study highlights the unique sequence and structural features of SecA protein specific to thermophilic and hyperthermophilic bacterial group. Our structural and computational analysis of these novel sequence features also provides some insights into the possible functions of one of these large CSIs in the context of thermostability of the protein. However, further understanding of the functional significance of these large CSIs in the functioning of this protein in thermophilic organisms and conferring thermostability will only emerge from future detailed genetic and biochemical studies.

Supplementary Materials: The following are available online at www.mdpi.com/. **Figure S1**: A Web logo representation of the sequence characteristics of the amino acids in the 50 aa and 70 aa CSIs. **Figure S2**: The root-mean-square-deviation (RMSD) values calculated relative to the starting structure for the *Thermotoga maritima* SecA (*Tm*SecA) with 50 insert (+CSI) and without 50 aa insert at 305.15K and 363.15K, over the 100 ns of molecular dynamics (MD) simulation trajectories. **Figure S3**: Trajectories for the occupancy of number of water molecules calculated by measuring the number of water molecules that are located with in the 9Å from ADP during the entire molecular dynamics (MD) simulation period of 100 ns. **Figure S4A and S4B:** Randomly picked snapshots at different time intervals extracted from the 100 ns MD trajectories of *Tm*SecA (+CSI) at 303.15K and 363.15K showing the occupancy and interaction of water molecules with residue (GLU185) from 50 aa CSI and ADP.

Author Contributions: BK and DP carried out identification of CSIs, phylogenetic analysis and homology modelling and localization of the CSIs in protein structures, and preparation of a draft manuscript; BK carried out the molecular dynamics (MD) simulation and analysis of the results. RSG, Planning, and supervision of the work, checking the identified CSIs for their specificities, obtained funding for the project and writing and finalizing of the manuscript.

Acknowledgments: This work was supported by Research Grant number 249924 from the Natural Science and Engineering Research Council of Canada awarded to Radhey S. Gupta.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Gaughran, E. R. The thermophilic microorganisms. *Bacteriol. Rev* 1947, **11**, 189-225.
2. Zeikus, J. G. Thermophilic bacteria: Ecology, physiology and technology. *Enzyme and Microbial Technology* 1979, **1**, 243-252.
3. Huber, R., Langworthy T.A., König, H., Thomm, T., Woese, C. R., Sleytr, U. B. & Stetter, K. O. *Thermotoga maritima* sp. nov. represents a new genus of unique extremely thermophilic eubacteria growing up to 90°C. *Archives of Microbiology* 1986, **144**, 324-333.
4. Dams, T., Auerbach, G., Bader, G., Jacob, U., Ploom, T., Huber, R. & Jaenicke, R. The crystal structure of dihydrofolate reductase from *Thermotoga maritima*: molecular features of thermostability. *J Mol. Biol.* 2000, **297**, 659-672.
5. Lee, D. W., Jang, H. J., Choe, E. A., Kim, B. C., Lee, S. J., Kim, S. B., Hong, Y. H. & Pyun, Y. R. Characterization of a thermostable L-arabinose (D-galactose) isomerase from the hyperthermophilic eubacterium *Thermotoga maritima*. *Appl. Environ. Microbiol.* 2004, **70**, 1397-1404.
6. Park, C. S., Yeom, S. J., Lim, Y. R., Kim, Y. S. & Oh, D. K. Characterization of a recombinant thermostable L: -rhamnose isomerase from *Thermotoga maritima* ATCC 43589 and its application in the production of L-lyxose and L-mannose. *Biotechnol. Lett.* 2010, **32**, 1947-1953.

7. Katrolia, P., Zhang, M., Yan, Q., Jiang, Z., Song, C. & Li, L. Characterisation of a thermostable family 42 â-galactosidase (BgalC) family from *Thermotoga maritima* showing efficient lactose hydrolysis. *Food Chemistry* 2011, **2**, 614-621.
8. Podar, M. & Reysenbach, A. L. New opportunities revealed by biotechnological explorations of extremophiles. *Curr. Opin. Biotechnol.* 2006, **17**, 250-255.
9. Robinson-Rechavi, M. & Godzik, A. Structural genomics of *Thermotoga maritima* proteins shows that contact order is a major determinant of protein thermostability. *Structure.* 2005, **13**, 857-860.
10. Wiegel, J., Ljungdhal, L. G. & Demain, A. L. The Importance of Thermophilic Bacteria in Biotechnology. *Critical Reviews in Biotechnology.* 1985, **3**, 39-108.
11. Lasa, I. & Berenguer, J. Thermophilic enzymes and their biotechnological potential. *Microbiologia* 1993, **9**, 77-89.
12. Turner, P., Mamo, G. & Karlsson, E. N. Potential and utilization of thermophiles and thermostable enzymes in biorefining. *Microb. Cell Fact.* 2007, **6**, 9.
13. Bhandari, V. & Gupta, R. S. Molecular signatures for the phylum (class) *Thermotogae* and a proposal for its division into three orders (*Thermotogales*, *Kosmotogales* ord. nov. and *Petrotogales* ord. nov.) containing four families (*Thermotogaceae*, *Fervidobacteriaceae* fam. nov., *Kosmotogaceae* fam. nov. and *Petrotogaceae* fam. nov.) and a new genus *Pseudothermotoga* gen. nov. with five new combinations. *Antonie Van Leeuwenhoek* 2014, **105**, 143-168.
14. Sadaf, A., Fatima, S. W. & Khare, S. K. (2019) in *Fungi in Extreme Environments: Ecological Role and Biotechnological Significance*, eds. Tiquia-Arashiro S. & Grube M. (Springer, Cham), pp. 307-328.
15. Mosina, N. L., Schubert, W. D. & Cowan, D. A. Characterization and homology modelling of a novel multi-modular and multi-functional *Paenibacillus mucilaginosus* glycoside hydrolase. *Extremophiles.* 2019, **23**, 681-686.
16. Vieille, C. & Zeikus, G. J. Hyperthermophilic enzymes: sources, uses, and molecular mechanisms for thermostability. *Microbiol. Mol. Biol. Rev.* 2001, **65**, 1-43.
17. Chien, A., Edgar, D. B. & Trela, J. M. Deoxyribonucleic acid polymerase from the extreme thermophile *Thermus aquaticus*. *J Bacteriol.* 1976, **127**, 1550-1557.
18. Sterner, R. & Liebl, W. Thermophilic adaptation of proteins. *Crit Rev. Biochem. Mol. Biol.* 2001, **36**, 39-106.
19. Zhou, X. X., Wang, Y. B., Pan, Y. J. & Li, W. F. Differences in amino acids composition and coupling patterns between mesophilic and thermophilic proteins. *Amino. Acids* 2008, **34**, 25-33.
20. Szilagyi, A. & Zavodszky, P. Structural differences between mesophilic, moderately thermophilic and extremely thermophilic protein subunits: results of a comprehensive survey. *Structure.* 2000, **8**, 493-504.
21. Russell, R. J., Ferguson, J. M., Hough, D. W., Danson, M. J. & Taylor, G. L. The crystal structure of citrate synthase from the hyperthermophilic archaeon *Pyrococcus furiosus* at 1.9 A resolution. *Biochemistry* 1997, **36**, 9983-9994.
22. Vogt, G., Woell, S. & Argos, P. Protein thermal stability, hydrogen bonds, and ion pairs. *J Mol. Biol* 1997, **269**, 631-643.
23. Bogin, O., Peretz, M., Hacham, Y., Korkhin, Y., Frolow, F., Kalb, G. & Burstein, Y. Enhanced thermal stability of *Clostridium beijerinckii* alcohol dehydrogenase after strategic substitution of amino acid residues with prolines from the homologous thermophilic *Thermoanaerobacter brockii* alcohol dehydrogenase. *Protein Sci* 1998, **7**, 1156-1163.
24. Russell, R. J., Gerike, U., Danson, M. J., Hough, D. W. & Taylor, G. L. Structural adaptations of the cold-active citrate synthase from an Antarctic bacterium. *Structure.* 1998, **6**, 351-361.
25. Glyakina, A. V., Garbuzynskiy, S. O., Lobanov, M. Y. & Galzitskaya, O. V. Different packing of external residues can explain differences in the thermostability of proteins from thermophilic and mesophilic organisms. *Bioinformatics* 2007, **23**, 2231-2238.
26. Le, P. T., Makhalanyane, T. P., Guerrero, L. D., Vikram, S., Van de, P. Y. & Cowan, D. A. Comparative Metagenomic Analysis Reveals Mechanisms for Stress Response in Hypoliths from Extreme Hyperarid Deserts. *Genome Biol. Evol* 2016, **8**, 2737-2747.
27. Gonzalez-Siso, M. I. Editorial for the Special Issue: Thermophiles and Thermozymes. *Microorganisms.* 2019, **7**.
28. Reysenbach, A.-L. (2001) in *Bergey's Manual of Systematic Bacteriology*, eds. Boone, D. R. & Castenholz, R. W. (Springer-Verlag, Berlin), pp. 359-367.

29. Reysenbach, A.-L. (2001) in *Bergey's Manual of Systematic Bacteriology*, eds. Boone, D. R. & Castenholz, R. W. (Springer-Verlag, Berlin), pp. 369-387.
30. Gupta, R. S. (2014) in *The Prokaryotes: Prokaryotic Biology and Symbiotic Associations.*, eds. Rosenberg E, DeLong, EF., Thompson, F., Lory, S., & Stackebrand, E. (Springer, Berlin), pp. 418-445.
31. Vaibhav Bhandari & Radhey S Gupta (2014) in *The Prokaryotes* (Springer Berlin Heidelberg, Berlin), pp. 989-1015.
32. Albuquerque, L. & Costa, M. S. (2014) in *The Prokaryotes- Other Major Lineages of Bacteria and the Archaea*, eds. Rosenberg, E., DeLong, E., Lory, S., Stackebrandt, E., & Thompson, F. (Springer, New York), pp. 955-987.
33. Escuder-Rodriguez, J. J., DeCastro, M. E., Cerdan, M. E., Rodriguez-Belmonte, E., Becerra, M. & Gonzalez-Siso, M. I. Cellulases from Thermophiles Found by Metagenomics. *Microorganisms.* 2018, **6**.
34. Bhandari, V. & Gupta, R. S. (2014) in *The Prokaryotes* (Springer Berlin Heidelberg, Berlin, Heidelberg), pp. 989-1015.
35. Rosenberg, E. (2014) in *The Prokaryotes- Other Major Lineages of Bacteria and the Archaea*, eds. Rosenberg, E., DeLong, E., Lory, S., Stackebrandt, E., & Thompson, F. (Springer, New York), pp. 613-615.
36. Huber, R. & Hannig, M. *Thermotogales. Prokaryotes* 2006, **7**, 899-922.
37. Pollo, S. M., Zhaxybayeva, O. & Nesbo, C. L. Insights into thermoadaptation and the evolution of mesophily from the bacterial phylum Thermotogae. *Can. J Microbiol* 2015, **61**, 655-670.
38. Griffiths, E. & Gupta, R. S. Distinctive protein signatures provide molecular markers and evidence for the monophyletic nature of the deinococcus-thermus phylum. *J Bacteriol.* 2004, **186**, 3097-3107.
39. Griffiths, E. & Gupta, R. S. Molecular signatures in protein sequences that are characteristics of the phylum Aquificae. *Int. J Syst. Evol Microbiol.* 2006, **56**, 99-107.
40. Gupta, R. S. & Bhandari, V. Phylogeny and molecular signatures for the phylum Thermotogae and its subgroups. *Antonie Van Leeuwenhoek* 2011, **100**, 1-34.
41. Gupta, R. S. & Lali, R. Molecular signatures for the phylum Aquificae and its different clades: proposal for division of the phylum Aquificae into the emended order *Aquificales*, containing the families *Aquificaceae* and *Hydrogenothermaceae*, and a new order *Desulfurobacteriales* ord. nov., containing the family *Desulfurobacteriaceae*. *Antonie Van Leeuwenhoek* 2013, **104**, 349-368.
42. Ho, J., Adeolu, M., Khadka, B. & Gupta, R. S. Identification of distinctive molecular traits that are characteristic of the phylum "Deinococcus-Thermus" and distinguish its main constituent groups. *Syst. Appl. Microbiol.* 2016, **39**, 453-463.
43. Rapoport, T. A., Jungnickel, B. & Kutay, U. Protein transport across the eukaryotic endoplasmic reticulum and bacterial inner membranes. *Annu. Rev. Biochem.* 1996, **65**, 271-303.
44. Fekkes, P. & Driessen, A. J. Protein targeting to the bacterial cytoplasmic membrane. *Microbiol. Mol. Biol. Rev.* 1999, **63**, 161-173.
45. Lill, R., Cunningham, K., Brundage, L. A., Ito, K., Oliver, D. & Wickner, W. SecA protein hydrolyzes ATP and is an essential component of the protein translocation ATPase of *Escherichia coli*. *EMBO J* 1989, **8**, 961-966.
46. Collinson, I. SecA-a New Twist in the Tale. *J Bacteriol.* 2017, **199**.
47. Chen, Y., Bauer, B. W., Rapoport, T. A. & Gumbart, J. C. Conformational Changes of the Clamp of the Protein Translocation ATPase SecA. *J Mol. Biol.* 2015, **427**, 2348-2359.
48. Vassylyev, D. G., Mori, H., Vassylyeva, M. N., Tsukazaki, T., Kimura, Y., Tahirov, T. H. & Ito, K. Crystal structure of the translocation ATPase SecA from *Thermus thermophilus* reveals a parallel, head-to-head dimer. *J Mol. Biol.* 2006, **364**, 248-258.
49. Zimmer, J. & Rapoport, T. A. Conformational flexibility and peptide interaction of the translocation ATPase SecA. *J Mol. Biol.* 2009, **394**, 606-612.
50. Griffiths, E. & Gupta, R. S. Signature sequences in diverse proteins provide evidence for the late divergence of the Order *Aquificales*. *Int. Microbiol.* 2004, **7**, 41-52.
51. Gupta, R. S. (2014) in *Methods in Microbiology New Approaches to Prokaryotics Systematics*, eds. Goodfellow M, Sutcliffe IC, & Chun J (Elsevier, London), pp. 153-182.
52. Larkin, M. A., Blackshields, G., Brown, N. P., Chenna, R., McGettigan, P. A., McWilliam, H., Valentin, F., Wallace, I. M., Wilm, A., Lopez, R. *et al.* Clustal W and Clustal X version 2.0. *Bioinformatics* 2007, **23**, 2947-2948.

53. Goujon, M., McWilliam, H., Li, W., Valentin, F., Squizzato, S., Paern, J. & Lopez, R. A new bioinformatics analysis tools framework at EMBL-EBI. *Nucleic Acids Res.* 2010, **38**, W695-W699.
54. Kumar, S., Nei, M., Dudley, J. & Tamura, K. MEGA: a biologist-centric software for evolutionary analysis of DNA and protein sequences. *Brief. Bioinform.* 2008, **9**, 299-306.
55. Whelan, S. & Goldman, N. A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. *Mol. Biol. Evol* 2001, **18**, 691-699.
56. Alnajar, S., Khadka, B. & Gupta, R. S. Ribonucleotide Reductases from Bifidobacteria Contain Multiple Conserved Indels Distinguishing Them from All Other Organisms: *In Silico* Analysis of the Possible Role of a 43 aa Bifidobacteria-Specific Insert in the Class III RNR Homolog. *Front Microbiol.* 2017, **8**, 1409.
57. Gupta, R. S., Nanda, A. & Khadka, B. Novel molecular, structural and evolutionary characteristics of the phosphoketolases from bifidobacteria and *Coriobacteriales*. *PLoS. One.* 2017, **12**, e0172176.
58. Khadka, B. & Gupta, R. S. Identification of a conserved 8 aa insert in the PIP5K protein in the *Saccharomycetaceae* family of fungi and the molecular dynamics simulations and structural analysis to investigate its potential functional role. *Proteins* 2017, **85**, 1454-1467.
59. Khadka, B., Adeolu, M., Blankenship, R. E. & Gupta, R. S. Novel insights into the origin and diversification of photosynthesis based on analyses of conserved indels in the core reaction center proteins. *Photosynth. Res.* 2017, **131**, 159-171.
60. Khadka, B. & Gupta, R. S. Novel Molecular Signatures in the PIP4K/PIP5K Family of Proteins Specific for Different Isozymes and Subfamilies Provide Important Insights into the Evolutionary Divergence of this Protein Family. *Genes* 2019, **10**, 312.
61. Shen, M. Y. & Sali, A. Statistical potential for assessment and prediction of protein structures. *Protein Sci.* 2006, **15**, 2507-2524.
62. Lee, G. R., Heo, L. & Seok, C. Effective protein model structure refinement by loop modeling and overall relaxation. *Proteins* 2016, **84 Suppl 1**, 293-301.
63. Lovell, S. C., Davis, I. W., Arendall, W. B., III, de Bakker, P. I., Word, J. M., Prisant, M. G., Richardson, J. S. & Richardson, D. C. Structure validation by Calpha geometry: phi,psi and Cbeta deviation. *Proteins* 2003, **50**, 437-450.
64. Colovos, C. & Yeates, T. O. Verification of protein structures: patterns of nonbonded atomic interactions. *Protein Sci.* 1993, **2**, 1511-1519.
65. Sippl, M. J. Recognition of errors in three-dimensional structures of proteins. *Proteins* 1993, **17**, 355-362.
66. Wiederstein, M. & Sippl, M. J. ProSA-web: interactive web service for the recognition of errors in three-dimensional structures of proteins. *Nucleic Acids Res.* 2007, **35**, W407-W410.
67. Eisenberg, D., Luthy, R. & Bowie, J. U. VERIFY3D: assessment of protein models with three-dimensional profiles. *Methods Enzymol.* 1997, **277**, 396-404.
68. Van Der, S. D., Lindahl, E., Hess, B., Groenhof, G., Mark, A. E. & Berendsen, H. J. GROMACS: fast, flexible, and free. *J Comput. Chem* 2005, **26**, 1701-1718.
69. Abrahama, M. J., Schulzb, R., Pálla, S., Smith, J. C., Hessa, B. & Lindahla, E. GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX* 2016, **1-2**, 19-25.
70. Jorgensen, W. L., Chandrasekhar, J., Madura, J. D., Impey, R. W. & Klein, M. L. Comparison of simple potential functions for simulating liquid water. *The Journal of Chemical Physics* 1983, **79**, 926-935.
71. Bussi, G., Donadio, D. & Parrinello, M. Canonical sampling through velocity rescaling. *J Chem Phys.* 2007, **126**, 014101.
72. Nose´, S. & M.L.Klein Constant pressure molecular dynamics for molecular systems. *Mol. Phys.* 1983, **50**, 1055-1076.
73. Parrinello, M. & A.Rahman Polymorphic transitions in single crystals: a new molecular dynamics method. *J. Appl. Phys.* 1981, **52**, 7182-7190.
74. Kutzner, C., Pall, S., Fechner, M., Esztermann, A., de Groot, B. L. & Grubmuller, H. Best bang for your buck: GPU nodes for GROMACS biomolecular simulations. *J Comput. Chem* 2015, **36**, 1990-2008.
75. Humphrey, W., Dalke, A. & Schulten, K. VMD: visual molecular dynamics. *J Mol. Graph.* 1996, **14**, 33-38.
76. Nesbo, C. L., Bradnan, D. M., Adebusuyi, A., Dlutek, M., Petrus, A. K., Foght, J., Doolittle, W. F. & Noll, K. M. *Mesotoga prima* gen. nov., sp. nov., the first described mesophilic species of the *Thermotogales*. *Extremophiles.* 2012, **16**, 387-393.

77. L'Haridon, S., Reysenbach, A. L., Tindall, B. J., Schonheit, P., Banta, A., Johnsen, U., Schumann, P., Gambacorta, A., Stackebrandt, E. & Jeanthon, C. *Desulfurobacterium atlanticum* sp. nov., *Desulfurobacterium pacificum* sp. nov. and *Thermovibrio guaymasensis* sp. nov., three thermophilic members of the *Desulfurobacteriaceae* fam. nov., a deep branching lineage within the Bacteria. *Int. J. Syst. Evol. Microbiol.* 2006, **56**, 2843-2852.
78. Hassan, F. M. N. & Gupta, R. S. Novel Sequence Features of DNA Repair Genes/Proteins from Deinococcus Species Implicated in Protection from Oxidatively Generated Damage. *Genes* 2018, **9**, 149.
79. Bonjour, F., Graber, A. & Aragno, M. Isolation of *Bacillus schlegelii*, a thermophilic, hydrogen oxidizing, aerobic autotroph, from geothermal and nongeothermal environments. *Microb. Ecol.* 1988, **16**, 331-337.
80. Kampfer, P., Glaeser, S. P. & Busse, H. J. Transfer of *Bacillus schlegelii* to a novel genus and proposal of *Hydrogenibacillus schlegelii* gen. nov., comb. nov. *Int. J Syst. Evol Microbiol.* 2013, **63**, 1723-1727.
81. Maker, A., Hemp, J., Pace, L. A., Ward, L. M. & Fischer, W. W. Draft Genome Sequence of *Hydrogenibacillus schlegelii* MA48, a Deep-Branching Member of the Bacilli Class of Firmicutes. *Genome Announc.* 2017, **5**.
82. Zhaxybayeva, O., Swithers, K. S., Lapierre, P., Fournier, G. P., Bickhart, D. M., Deboy, R. T., Nelson, K. E., Nesbo, C. L., Doolittle, W. F., Gogarten, J. P. *et al.* On the chimeric nature, thermophilic origin, and phylogenetic placement of the *Thermotogales*. *Proc. Natl. Acad. Sci. U. S A* 2009, **106**, 5865-5870.
83. Hunt, J. F., Weinkauf, S., Henry, L., Fak, J. J., McNicholas, P., Oliver, D. B. & Deisenhofer, J. Nucleotide control of interdomain interactions in the conformational reaction cycle of SecA. *Science* 2002, **297**, 2018-2026.
84. Papanikolau, Y., Papadovasilaki, M., Ravelli, R. B., McCarthy, A. A., Cusack, S., Economou, A. & Petratos, K. Structure of dimeric SecA, the *Escherichia coli* preprotein translocase motor. *J Mol. Biol.* 2007, **366**, 1545-1557.
85. Sharma, V., Arockiasamy, A., Ronning, D. R., Savva, C. G., Holzenburg, A., Braunstein, M., Jacobs, W. R., Jr. & Sacchettini, J. C. Crystal structure of *Mycobacterium tuberculosis* SecA, a preprotein translocating ATPase. *Proc. Natl. Acad. Sci. U. S. A* 2003, **100**, 2243-2248.
86. Vrontou, E. & Economou, A. Structure and function of SecA, the preprotein translocase nanomotor. *Biochim Biophys Acta* 2004, **1694**, 67-80.
87. Zimmer, J., Nam, Y. & Rapoport, T. A. Structure of a complex of the ATPase SecA and the protein-translocation channel. *Nature* 2008, **455**, 936-943.
88. Merritt, E. A., Sixma, T. K., Kalk, K. H., van Zanten, B. A. & Hol, W. G. Galactose-binding site in *Escherichia coli* heat-labile enterotoxin (LT) and cholera toxin (CT). *Mol. Microbiol.* 1994, **13**, 745-753.
89. Barillari, C., Taylor, J., Viner, R. & Essex, J. W. Classification of water molecules in protein binding sites. *J Am. Chem Soc* 2007, **129**, 2577-2587.
90. Lu, Y., Wang, R., Yang, C. Y. & Wang, S. Analysis of ligand-bound water molecules in high-resolution crystal structures of protein-ligand complexes. *J Chem Inf. Model.* 2007, **47**, 668-675.
91. Homans, S. W. Water, water everywhere--except where it matters? *Drug Discov. Today* 2007, **12**, 534-539.
92. Singh, N. & Briggs, J. M. Molecular dynamics simulations of Factor Xa: insight into conformational transition of its binding subsites. *Biopolymers* 2008, **89**, 1104-1113.
93. Schiebel, J., Gaspari, R., Wulsdorf, T., Ngo, K., Sohn, C., Schrader, T. E., Cavalli, A., Ostermann, A., Heine, A. & Klebe, G. Intriguing role of water in protein-ligand binding studied by neutron crystallography on trypsin complexes. *Nat. Commun.* 2018, **9**, 3559.
94. Vieille, C., Burdette, D. S. & Zeikus, J. G. Thermozymes. *Biotechnol. Annu Rev* 1996, **2**, 1-83.
95. Kumar, S., Tsai, C. J. & Nussinov, R. Factors enhancing protein thermostability. *Protein Eng* 2000, **13**, 179-191.
96. Berezovsky, I. N. & Shakhnovich, E. I. Physics and evolution of thermophilic adaptation. *Proc Natl. Acad. Sci U. S. A* 2005, **102**, 12742-12747.
97. Mizuguchi, K., Sele, M. & Cubellis, M. V. Environment specific substitution tables for thermophilic proteins. *BMC. Bioinformatics* 2007, **8 Suppl 1**, S15.
98. Taylor, T. J. & Vaisman, I. I. Discrimination of thermophilic and mesophilic proteins. *BMC. Struct. Biol* 2010, **10 Suppl 1**, S5.
99. Singh, B. & Gupta, R. S. Conserved inserts in the Hsp60 (GroEL) and Hsp70 (DnaK) proteins are essential for cellular growth. *Mol. Genet. Genomics* 2009, **281**, 361-373.

100. Huber, R. & Eder, W. *Aquificales. Prokaryotes* 2006, **7**, 938.
101. Bonch-Osmolovskaya, E. *Aquificales.* 2008 Wieley, pp. 1-7.
102. Woese, C. R. Bacterial evolution. *Microbiol. Rev.* 1987, **51**, 221-271.
103. Knight, J. D., Hamelberg, D., McCammon, J. A. & Kothary, R. The role of conserved water molecules in the catalytic domain of protein kinases. *Proteins* 2009, **76**, 527-535.
104. Nygaard, R., Valentin-Hansen, L., Mokrosinski, J., Frimurer, T. M. & Schwartz, T. W. Conserved water-mediated hydrogen bond network between TM-I, -II, -VI, and -VII in 7TM receptor activation. *J Biol Chem* 2010, **285**, 19625-19636.
105. Kaur, M., Bahia, M. S. & Silakari, O. Exploring the role of water molecules for docking and receptor guided 3D-QSAR analysis of naphthyridine derivatives as spleen tyrosine kinase (Syk) inhibitors. *J Chem Inf. Model.* 2012, **52**, 2619-2630.
106. Jeszenoi, N., Balint, M., Horvath, I., Van Der, S. D. & Hetenyi, C. Exploration of Interfacial Hydration Networks of Target-Ligand Complexes. *J Chem Inf. Model.* 2016, **56**, 148-158.
107. Lett, C. M., Berghuis, A. M., Frey, H. E., Lepock, J. R. & Guillemette, J. G. The role of a conserved water molecule in the redox-dependent thermal stability of iso-1-cytochrome c. *J Biol Chem* 1996, **271**, 29088-29093.
108. Shi, R. & Lin, S. X. Cofactor hydrogen bonding onto the protein main chain is conserved in the short chain dehydrogenase/reductase family and contributes to nicotinamide orientation. *J Biol Chem* 2004, **279**, 16778-16785.
109. Sterpone, F., Stirnemann, G., Hynes, J. T. & Laage, D. Water hydrogen-bond dynamics around amino acids: the key role of hydrophilic hydrogen-bond acceptor groups. *J Phys Chem B* 2010, **114**, 2083-2089.
110. Huggins, D. J. & Tidor, B. Systematic placement of structural water molecules for improved scoring of protein-ligand interactions. *Protein Eng Des Sel* 2011, **24**, 777-789.
111. Milenkovic, S. & Bondar, A. N. Mechanism of conformational coupling in SecA: Key role of hydrogen-bonding networks and water interactions. *Biochim Biophys Acta* 2016, **1858**, 374-385.
112. Geszvain, K., Gruber, T. M., Mooney, R. A., Gross, C. A. & Landick, R. A hydrophobic patch on the flap-tip helix of *E.coli* RNA polymerase mediates sigma(70) region 4 function. *J. Mol. Biol.* 2004, **343**, 569-587.
113. Akiva, E., Itzhaki, Z. & Margalit, H. Built-in loops allow versatility in domain-domain interactions: lessons from self-interacting domains. *Proc. Natl. Acad. Sci. U. S. A* 2008, **105**, 13292-13297.
114. Hashimoto, K. & Panchenko, A. R. Mechanisms of protein oligomerization, the critical role of insertions and deletions in maintaining different oligomeric states. *Proc. Natl. Acad. Sci. U. S. A* 2010, **107**, 20352-20357.
115. Schoeffler, A. J., May, A. P. & Berger, J. M. A domain insertion in *Escherichia coli* GyrB adopts a novel fold that plays a critical role in gyrase function. *Nucleic Acids Res.* 2010, **38**, 7830-7844.
116. Clarke, J. H. & Irvine, R. F. Evolutionarily conserved structural changes in phosphatidylinositol 5-phosphate 4-kinase (PI5P4K) isoforms are responsible for differences in enzyme activity and localization. *Biochem. J* 2013, **454**, 49-57.
117. Gouridis, G., Karamanou, S., Sardis, M. F., Scharer, M. A., Capitani, G. & Economou, A. Quaternary dynamics of the SecA motor drive translocase catalysis. *Mol. Cell* 2013, **52**, 655-666.
118. Singh, R., Kraft, C., Jaiswal, R., Sejwal, K., Kasaragod, V. B., Kuper, J., Burger, J., Mielke, T., Luirink, J. & Bhushan, S. Cryo-electron microscopic structure of SecA protein bound to the 70S ribosome. *J Biol. Chem* 2014, **289**, 7190-7199.
119. Wowor, A. J., Yan, Y., Auclair, S. M., Yu, D., Zhang, J., May, E. R., Gross, M. L., Kendall, D. A. & Cole, J. L. Analysis of SecA dimerization in solution. *Biochemistry* 2014, **53**, 3248-3260.
120. Collinson, I., Corey, R. A. & Allen, W. J. Channel crossing: how are proteins shipped across the bacterial plasma membrane? *Philos. Trans. R. Soc Lond B Biol. Sci.* 2015, **370**.
121. Fekkes, P., de Wit, J. G., van der Wolk, J. P., Kimsey, H. H., Kumamoto, C. A. & Driessen, A. J. Preprotein transfer to the *Escherichia coli* translocase requires the co-operative binding of SecB and the signal sequence to SecA. *Mol. Microbiol.* 1998, **29**, 1179-1190.
122. Pretz, M. G., Remigy, H., Swaving, J., Albers, S. V., Garrido, V. G., Chami, M., Engel, A. & Driessen, A. J. Functional and structural characterization of the minimal Sec translocase of the hyperthermophile *Thermotoga maritima*. *Extremophiles.* 2005, **9**, 307-316.

123. Sala, A., Calderon, V., Bordes, P. & Genevaux, P. TAC from *Mycobacterium tuberculosis*: a paradigm for stress-responsive toxin-antitoxin systems controlled by SecB-like chaperones. *Cell Stress. Chaperones.* 2013, **18**, 129-135.

SUPPLEMENTARY INFORMATION

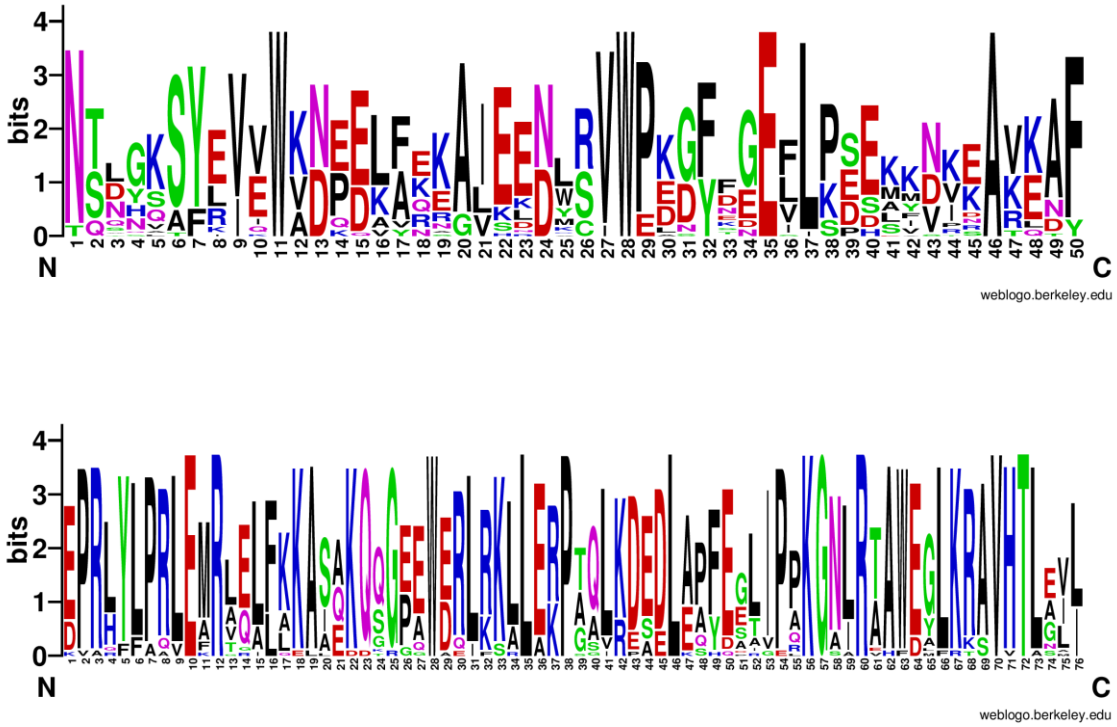


Figure S1. A Web logo representation of the sequence characteristics of amino acids in **(A)** 50 aa CSI and **(B)** 70 aa CSI. Web logos were created by using the CSIs entered into the Weblogo program (http://weblogo.berkeley.edu/) using default parameters. Amino acids are color-coded according to their chemical properties as polar amino acids (G,S,T,Y,C,Q,N) are green, basic (K, R, H) are blue, acidic (D,E) are red and hydrophobic (A,V,L,I,P,W,F,M) are black.

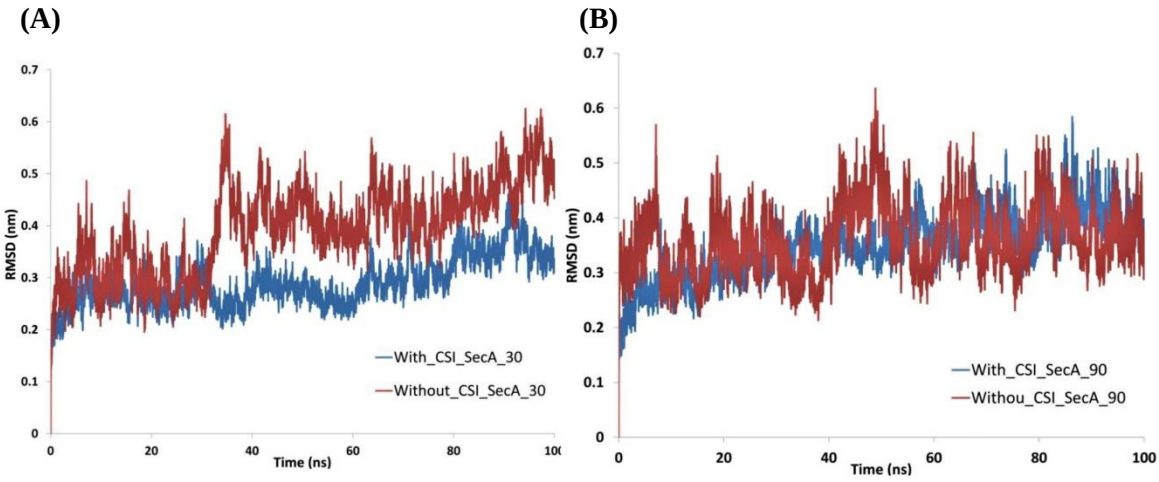


Figure S2. The root-mean-square-deviation (RMSD) values calculated relative to the starting structure for the *Thermotoga maritima* SecA (*Tm*SecA) with 50 insert (+CSI) and without 50 aa insert at (A) 305.15K and (B) 363.15K, over the 100 ns of molecular dynamics (MD) simulation trajectories.

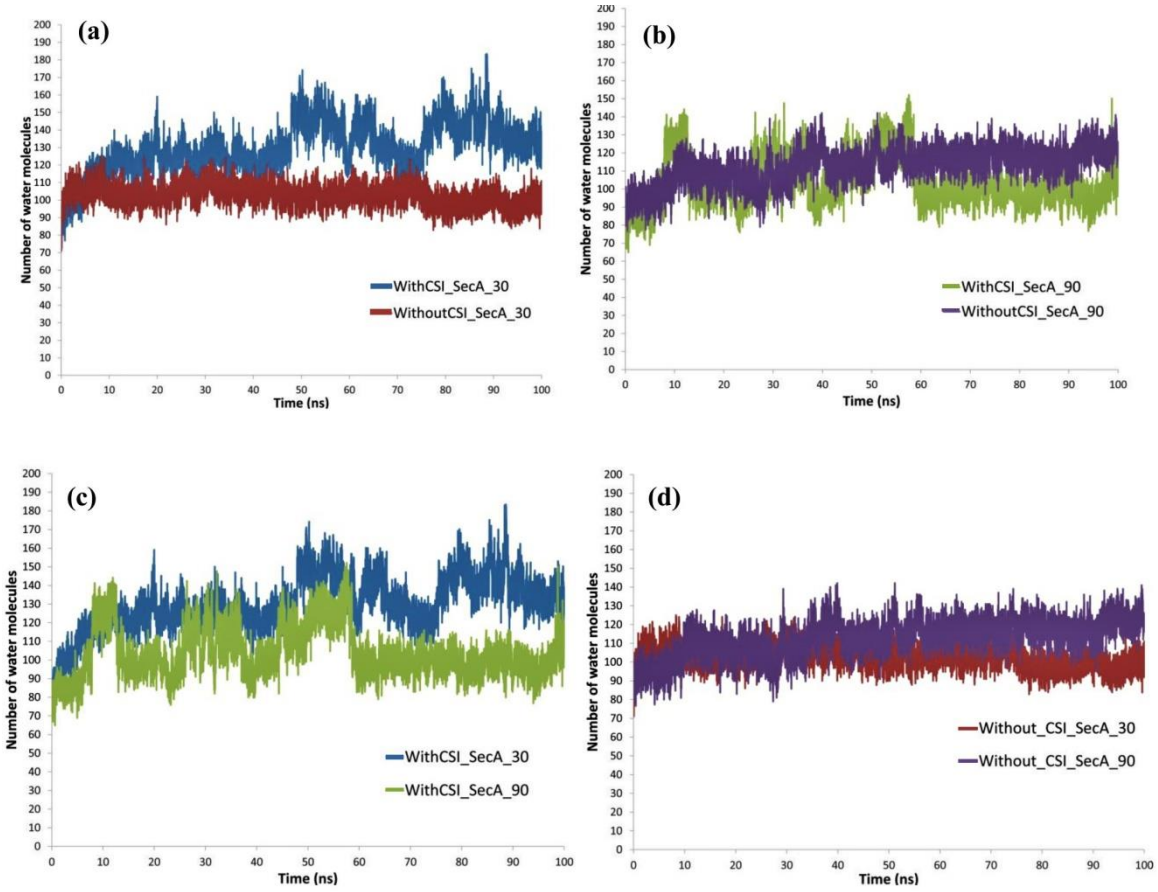


Figure S3. Trajectories for the occupancy of the number of water molecules calculated by measuring the number of water molecules that are located within the 9Å from ADP during the entire molecular dynamics (MD) simulation period of 100 ns. Occupancy of number of water molecules compared between the *Tm*SecA with (+CSI) and without (-CSI) 50 aa insertion at **(a)** 305.15K and **(b)** 363.15K. Comparisons of the number of water molecules calculated for **(c)** *Tm*SecA (+CSI) and **(d)** *Tm*SecA (-CSI) at two different temperatures as labeled in the plots.

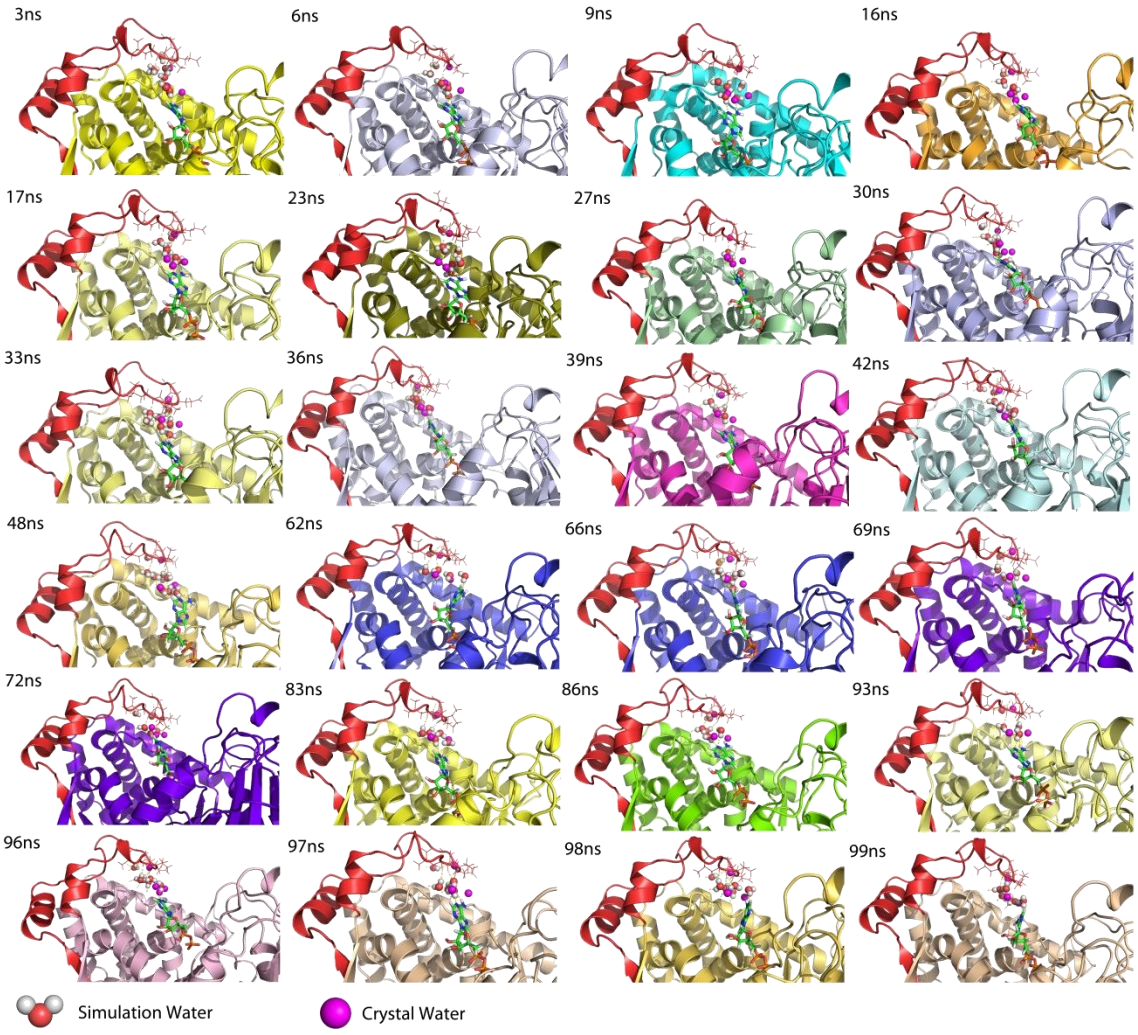


Figure S4 (A). Randomly picked snapshots at different time intervals extracted from the 100 ns MD trajectories of *Tm*SecA (+CSI) at 303.15K showing the occupancy and interaction of water molecules with residue (GLU185) from 50 aa CSI and ADP.

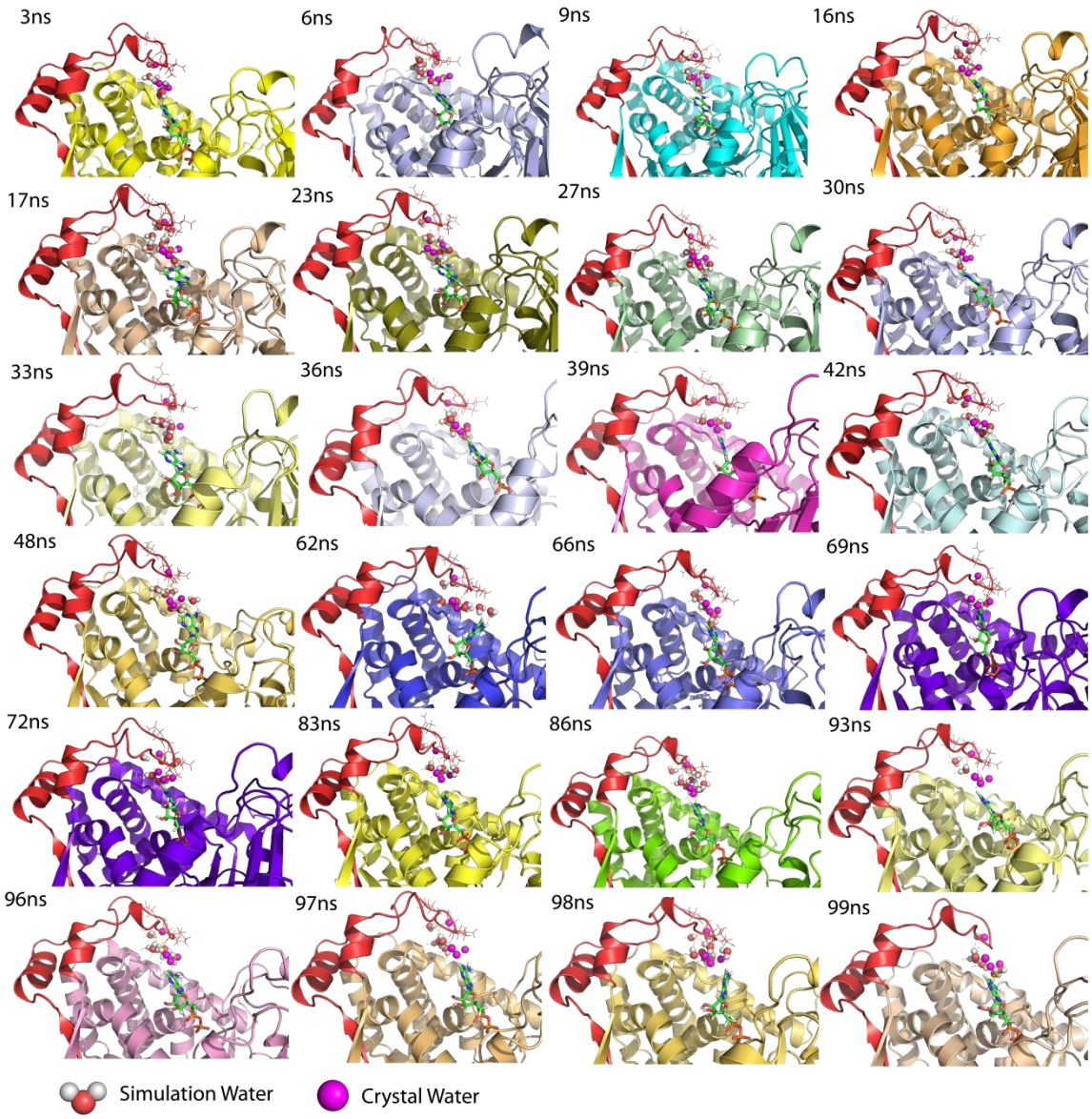


Figure S4 (B). Randomly picked snapshots at different time intervals extracted from the 100 ns MD trajectories of TmSecA (+CSI) at 363.15K showing the occupancy and interaction of water molecules with residue (GLU185) from 50 aa CSI and ADP.

CHAPTER 6

GlabModeller: A Graphical User Interface to a Streamed line Pipeline for Homology Modelling Process.

Background

Knowledge of the three-dimensional (3D) structure of the protein is crucial for understanding of its basic function (Sippl, 1993). The advent of the genome sequencing project have resulted in the availability of enormous wealth of known protein sequence data (McPherson, 2014). At the same time, the numbers of experimentally solved protein structures are lagging increasingly behind, owing to the difficulties associated with the determination of protein structure. As a result, the gap between the number of sequence information and corresponding structural information is widening rapidly (Mistry et al., 2013). As of July 2019, the number of structures deposited in the freely accessible online database called the Protein Data Bank (PDB) has reached 154,015 structures (Rose et al., 2013; Burley et al., 2019). Yet, this data looks very tiny in comparison to the astonishing number of sequence data held by the Uniprot Knowledge database (Uniprot), consisting information of about more than 120 million sequences, as of Uniprot release 2019 (UniProt Consortium, 2019). Most of the structures of the proteins deposited in the PDB are solved by X-Ray Crystallography (Burley et al., 2019), which is considered the most powerful method for obtaining the 3D structures in atomic details (Ilari and Savino, 2008). Despite successful applications, it still has intrinsic limitations such as significant investment of efforts and experimental time for purification and inability to provide information about the dynamic behavior of proteins (Ostermeier and Michel, 1997; Henzler-Wildman and Kern, 2007; Davis et al., 2008). In contrast to X-ray crystallography, Nuclear Magnetic Resonance (NMR) method can provide observations about the dynamic of the protein in solution. However, this method is limited to

comparatively smaller size proteins (Mittermaier and Kay, 2006; Boehr et al., 2009; Kleckner and Foster, 2011). On the other hand, Cryo-Electron Microscopy (Cryo-EM), which is an essential complement to both X-ray Crystallography and NMR, extends its capability in determination of the 3D structure of macromolecules. Cryo-EM has continued to be a powerful tool due to its ability to confront colossal assemblies and transient, and requirement of small amount of starting materials. Moreover, it is well suited to study membrane protein within native membranous lipid environment. However, this method is particularly limited to the low level of atomic resolution (Saibil, 2000; Ubarretxena-Belandia and Stokes, 2010). In the absence of an experimentally determined atomic-resolution protein structures, computational methods such as homology modelling are playing an increasingly important role to bridge the gap created by rapidly growing sequenced protein sequence universe and the world of solved protein structures (Mistry et al., 2013; Kryshtafovych et al., 2014; McPherson, 2014). Homology modelling allows the generation of 3D models of a (target) protein from its primary amino acid sequences based on its alignment with one or more homologous proteins with known 3D structure (template) that share statistically significant sequence similarity (Chothia and Lesk, 1986; Sali and Blundell, 1993; Ginalski, 2006). In general, the workflow of homology or comparative modelling can be divided into following basic steps (Figure 6.1):

1. Searching for known 3D structures that could serve as a suitable template for a given target protein sequence; sequence identity of >30% between target and template is generally considered threshold for successful homology modelling (Fiser, 2010; Kryshtafovych et al., 2014).

2. Alignment of target sequence with template; may include manual adjustment of multiple sequence alignment to optimize the placement of Insertions/Deletions (Indels) outside tight secondary structure elements (Pascarella and Argos, 1992).
3. 3D model construction based on information of alignment and template structure (Sali and Blundell, 1993).
4. Refinement and assessment of the resulting model; represents an essential component of protein modelling, as the accuracy/quality of final model will determine its usefulness for various applications (Sanchez and Sali, 1997; Marti-Renom et al., 2000; Baker and Sali, 2001; Xu and Zhang, 2012).

The protein model generated using computational methods (e.g. Homology modelling) can provide a wide range of useful applications to molecular biology, for generating and testing several hypothesis such as: (i) substrate specificity (Lewis et al., 1999; Lewis, 1999; De Rienzo et al., 2000; Lukk et al., 2012), (ii) rational design of mutagenesis experiments (Chmiel et al., 2005; de Graaf et al., 2007), (iii) predicting and analyzing ligand binding sites and ligand-receptor interactions (Chen et al., 2005; Carlsson et al., 2011; Levit et al., 2012; Nguyen et al., 2016), (iv) rational drug design (Hillisch et al., 2004; Thiel, 2004; Park et al., 2008; Sharma et al., 2012; Schmidt et al., 2014), (v) as a starting model for solving structure from X-ray crystallography, Electron Microscopy and NMR (Sutcliffe et al., 1992; Ceulemans and Russell, 2004; Ngo et al., 2008), (vi) identifying and characterizing the protein-protein complexes (Launay and Simonson, 2008; Kundrotas et al., 2012), and (vii) carrying out Molecular dynamics (MD) simulation studies of biological macromolecules to gain insight into the physical basis of

structure and function (Capener et al., 2000; Karplus, 2002; Karplus and Kuriyan, 2005; Sahoo et al., 2014). One of the most widely regarded computational tools for protein structure prediction is MODELLER (Sali and Blundell, 1993). The lack of a Graphical User Interference (GUI) for MODELLER requires the user to have detailed knowledge about its manual and other tutorials, and also a basic understanding of python scripting to perform different tasks. In general, the process of model generation using the comparative modelling tool, MODELLER, usually requires, compiling and setting up of input files, and editing of python scripts for different steps, which can be very time consuming when dealing with a large number of target proteins. In this regard, we have created an interactive graphical pipeline program for protein modelling process known as "*Gupta Lab-Modeller* or *GlabModeller*". GlabModeller is a standalone program, which provides an easy-to-use graphical user interface (GUI) to MODELLER and a number of subsequent steps in the model refinement and the model validation process. In addition, several other features such as template selection, multiple sequence alignment (MSA) editing and manual specification of spatial restraint have now also been implemented. Several interfaces for MODELLER or Homology Modelling process, mostly in the form of automated servers, have been described previously (Schwede et al., 2003; Pettersen et al., 2004; Kelley et al., 2015). While these tools, in some cases, are suitable for challenging targets (target that shares low sequence identity with template) (Kopp and Schwede, 2004; Yang and Zhang, 2015), they are oftentimes very time consuming and limit the user from more critical analysis of the generated models. Additionally, they lack the features such as the execution of spatial restraint and accesses to the streamlined

validation analysis using multiple tools are limited. GlabModeller, which provides a seamless interface to MODELLER, should enhance the overall protein modelling process by simplifying and speeding up the key modelling steps without the user requiring any knowledge of the backend applications.

Graphical User Interface (GUI) of GlabModeller

Shown in Figure 6.2 is the easy to run GUI of the GlabModeller pipeline written using the python Tk interface module. The GUI for GlabModeller comprises of three major components: the sequence-template alignment, the model building and refinement, and the model validation. The "sequence-template" section allows the user to use a raw amino acid sequence of the target protein and to select the directory containing template protein files in the PDB format. In the "*Model-building and refinement*" section, the user can select the aligned files (generated from the "sequence-template" section), select the directory containing the templates, and select the directory folder in which the generated output protein models will be stored. This section also allows the user to adjust the number of models to be built and to input other variables such as defining secondary structure information to be restrained in the model. Additionally, the user has option to run energy minimization of the generated top ten models with high DOPE score by checking the "*Run ModRefiner*" box. The ModRefiner program allows atomic-level high-resolution structural refinement to remove the steric clashes and to improve hydrogen bonding network, side-chain positioning and backbone topology of the generated models (Xu and Zhang, 2011). The program can also be accessed through an online server at the Zang Lab website at https://zhanglab.ccmb.med.umich.edu/ModRefiner.

The final section, the "*Model validation"*, constitute a critical step in the model building process as it helps to identify potential errors in the predicted structural models. This section provides an option to run the four independent validation tools viz. RAMPAGE (Lovell et al., 2003), ProSA (Sippl, 1993; Wiederstein and Sippl, 2007), ERRAT (Colovos and Yeates, 1993) and VERIFY3D (Bowie et al., 1991; Luthy et al., 1992; Eisenberg et al., 1997). RAMPAGE is a structure validation tool for the assessment of Ramachandran plots which allow the visualization of energetically favored and disallowed dihedral angles psi (ψ) and phi (φ) calculated based on van der radius of their side chains (Kleywegt and Jones, 1996; Carrascoza et al., 2014). The results from RAMPAGE include number/percentage of residues in the favored region, allowed region and outlier region, and allow to access the stereochemical quality of the generated models. The ProSA (Protein Structure Analysis) program is a well-known and widely used tool which is frequently employed in the refinement and validation of the experimentally derived protein structures or theoretical models obtained from homology modeling (Wiederstein and Sippl, 2007). The Z-score predicted using ProSA web server provides an indication of the overall quality of a model and it is commonly used to ensure the compatibility of the Z-scores range between input target protein structures and native protein of similar size (Sippl, 1993; Wiederstein and Sippl, 2007). The ERRAT program analyzes the statistics of non-bonded interactions between various atom types and the value of error functions versus position of nine residues sliding window are plotted by comparing with the statics from a database of reliable and high resolution crystallography structures. The ERRAT quality factor value is expressed as the percentage of the protein

residues in which the calculated error value falls below the 95% rejection limit. A high resolution crystal structures generally show an ERRAT quality factor value of 95% or higher, whereas a low resolution structure generally shows a value of approximately around 80% (Colovos and Yeates, 1993). VERIFY3D program determines the compatibility of an atomic model of 3D structures with its own primary amino acid sequence (1D) by assigning a structural class based on its location and environment (alpha, beta, loop, polar, nonpolar, and other properties) and by comparing the results to reliable structures (Eisenberg et al., 1997). All the results and output files obtained from the aforementioned validation tools can be stored in the user defined output folder.

Other Requirements to run GlabModeller and analysis of the results

GlabModeller requires the recent version of MODELLER and Python installed in the local system in the default home directory. MODELLER is available free of charge to academic non-profit institutions. Information regarding links to register for an academic license key and to download and install the different versions of MODELLER compatible with various operating systems can be access through its home page at the Sali Lab website (www.salilab.org). The 3D coordinates of the generated homology models can be visualized and analyzed using molecular visualization tools such as PyMOL (www.pymol.org) and Cn3D ("see in 3D") (Wang et al., 2000). Secondary elements information can be identified by cartoon representation/rendering with arrows (β-strands), spirals (α-helices) and lines (loops/turns) to simplify the depiction of protein structure architecture. Additionally, any region of interest (e.g. CSIs) can be color-coded to make

their interpretation easier. Currently, GlabModeller only runs on Microsoft Windows. In future, the binary executables and source code for the GlabModeller will be hosted on GitHub.

Mapping the CSIs in DNA-dependent RNA polymerase Alpha Subunit (RpoA) and DNA-dependent RNA polymerase Beta Subunit (RpoB) using GlabModeller

DNA-dependent RNA polymerase Alpha Subunit (RpoA) and DNA-dependent RNA polymerase Beta Subunit (RpoB) comprises the two major components of the core RNA polymerase enzyme complex (Ebright and Busby, 1995; Severinov et al., 1996; Minakhin et al., 2001). The RpoA and RpoB subunit together with other large subunits of RNA polymerase complex form the catalytic center of the enzyme and a binding site for double-helix DNA, nucleotide substrates and nascent RNA (Landick et al., 1990; Mustaev et al., 1993; Gross et al., 1996). Antibiotics like Rifamycins and its structural counterparts that are known to inhibit RNA polymerase activity acts by binding to the beta subunit of RNA polymerase (Ho et al., 2016) and the majority of Rifamycins resistances in bacterial mutants have also been mapped into the beta subunit of RNA polymerase (Campbell et al., 2001; Ho et al., 2016). Earlier studies from our lab have identified 3 different CSIs (3 insertions) in RpoA and 11 different CSIs (8 insertions and 3 deletions) in the RpoB protein, which are specific for a different group of microbes (Table 1). Although, the experimentally solved structural information is available for some of the CSIs, such as in the case of widely studied large prominent signature inserts of >100 aa, specific for Proteobacteria, the CFBG group, Chlamydiae, and Aquificales (Gupta, 2004). However, the structural information for most other CSI containing RpoA

and RpoB is not available. In view of this, I have utilized the GlabModeller to create the homology model to analyze the structural features of the CSIs in these proteins specific for different groups of microbes. In brief, the PSI-BLAST was carried out to identify the suitable templates for each target protein (Table 1). Initially, 200 structures were generated for each target (RpoA and RpoB) proteins and selected using DOPE score (Shen and Sali, 2006) and are then submitted to the ModRefiner program to obtain atomic-level energy minimization and to obtain a model with reliable stereochemistry quality (Xu and Zhang, 2012). The validation of the refined model was carried out using the tools available in GlabModeller. The results for the stereochemical assessment of the RpoA and RpoB homology models are shown in Table 2. Overall, analysis of the structural features of the identified CSIs in RpoA and RpoB, using the homology models generated using GlabModeller, shows that all identified CSIs are found to be located on the surface loop of proteins, and they are generally situated away from the active site without disrupting the core function of a protein. In RpoB, the majority of CSIs were found to be clustered in the beta-2 domain (in *E. coli* residues ranges from 151-445 aa) and flap domain (in *E. coli* residues ranges from 831-1057) of RpoB. The flexible flab domain are reported to involved in the interaction with sigma factors for promoter recognition activities (Kuznedelov et al., 2002; Geszvain et al., 2004) whereas the beta-2 domain forms a part of the inner roof for downstream double-strand DNA binding channel (Korzheva et al., 2000; Lane and Darst, 2010).

Discussion

GlabModeller is a streamlined pipeline that provides a user-friendly interface to MODELLER and a number of subsequent steps involved in the homology modelling process. It was developed with an aim to assist an efficient analysis of large number CSIs that are identified in various essential proteins specific for a different group of organisms. The previous version of this pipeline has already been utilized to carry out the study on mapping of the structural location of various CSIs identified in number of different functionally important proteins such as DNA-dependent RNA polymerase B (RpoB), Ribonucleotide reductase (RNA), photosynthetic reaction core proteins etc. (Gupta and Khadka, 2015; Alnajar et al., 2017; Khadka et al., 2017; Hassan and Gupta, 2018; Khadka et al., 2019). The GlabModeller was also utilized for the generation of homology models and the analyses of the structural features of identified CSIs in PIP4K/PIP5K family of proteins (Khadka and Gupta, 2017; Khadka and Gupta, 2019), as described in the work in the Chapter 2, and Chapter 3 of this thesis, and as well as phosphoketolase (PK) (Gupta et al., 2017) and SecA proteins as described in Chapter 4 and Chapter 5 of this thesis. In addition, a list of my other published articles in which the GlabModeller was utilized to generate the homolog models and analyses of the structural location of CSIs in protein structure is shown in Table 3. Overall, our integrated software pipeline for an automated and efficient generation of homology models GlabModeller pipeline will make the homology modelling approach available to users from broader scientific communities. Specifically, this will greatly benefit novice users and experimentalists to overcome the initial learning curves barrier hindering the general use of homology modelling approach.

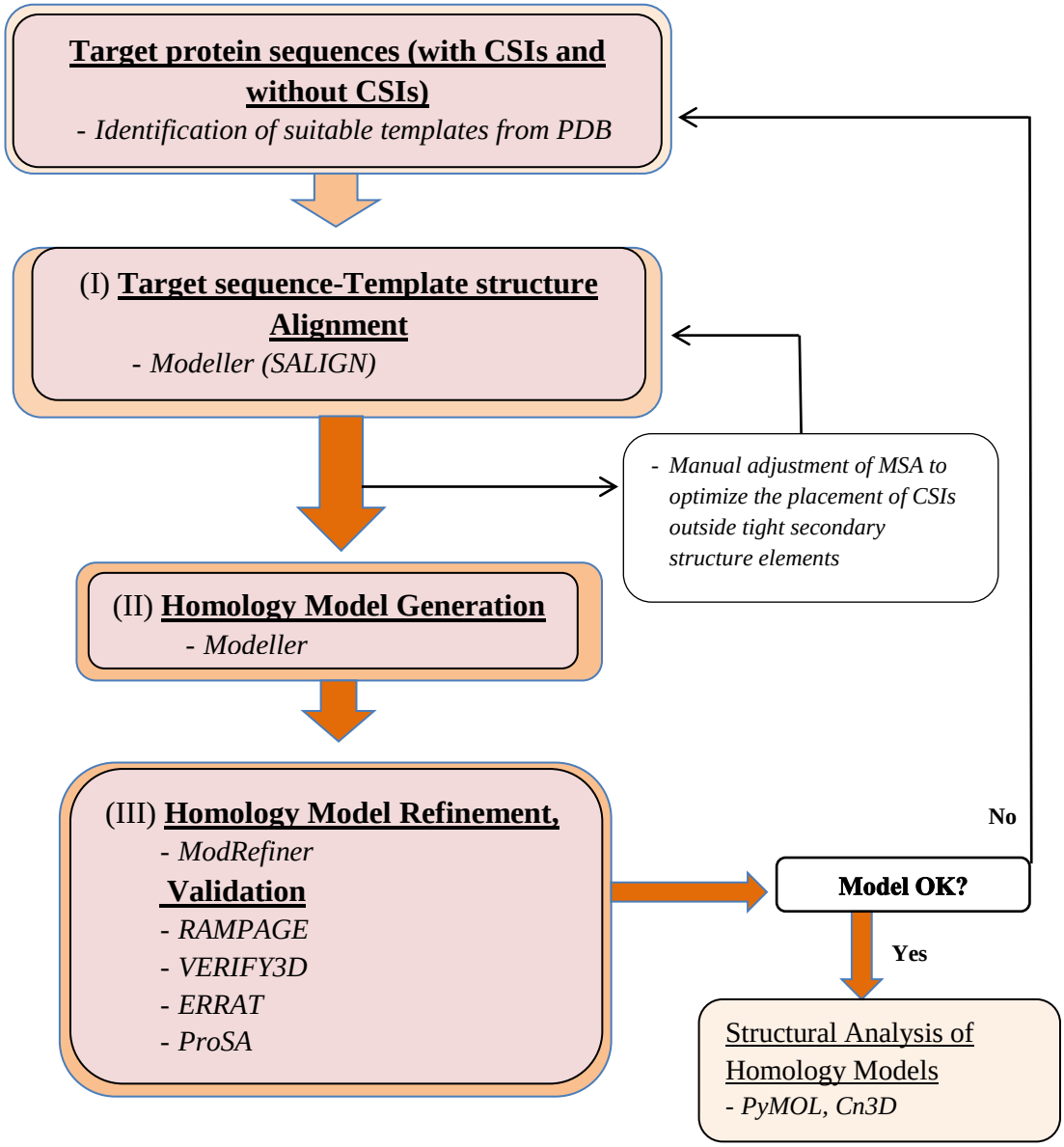


Figure 6.1. Workflow depicting the pipeline protocol to prepare run and analyze homology modelling. After the suitable template structure has been identified, the core of the pipeline consists of three major sections: (I) target sequence-template alignment; (II) homology model generation; (III) model refinement, validation, and analysis. The name of the programs utilized in each step of the pipeline is indicated.

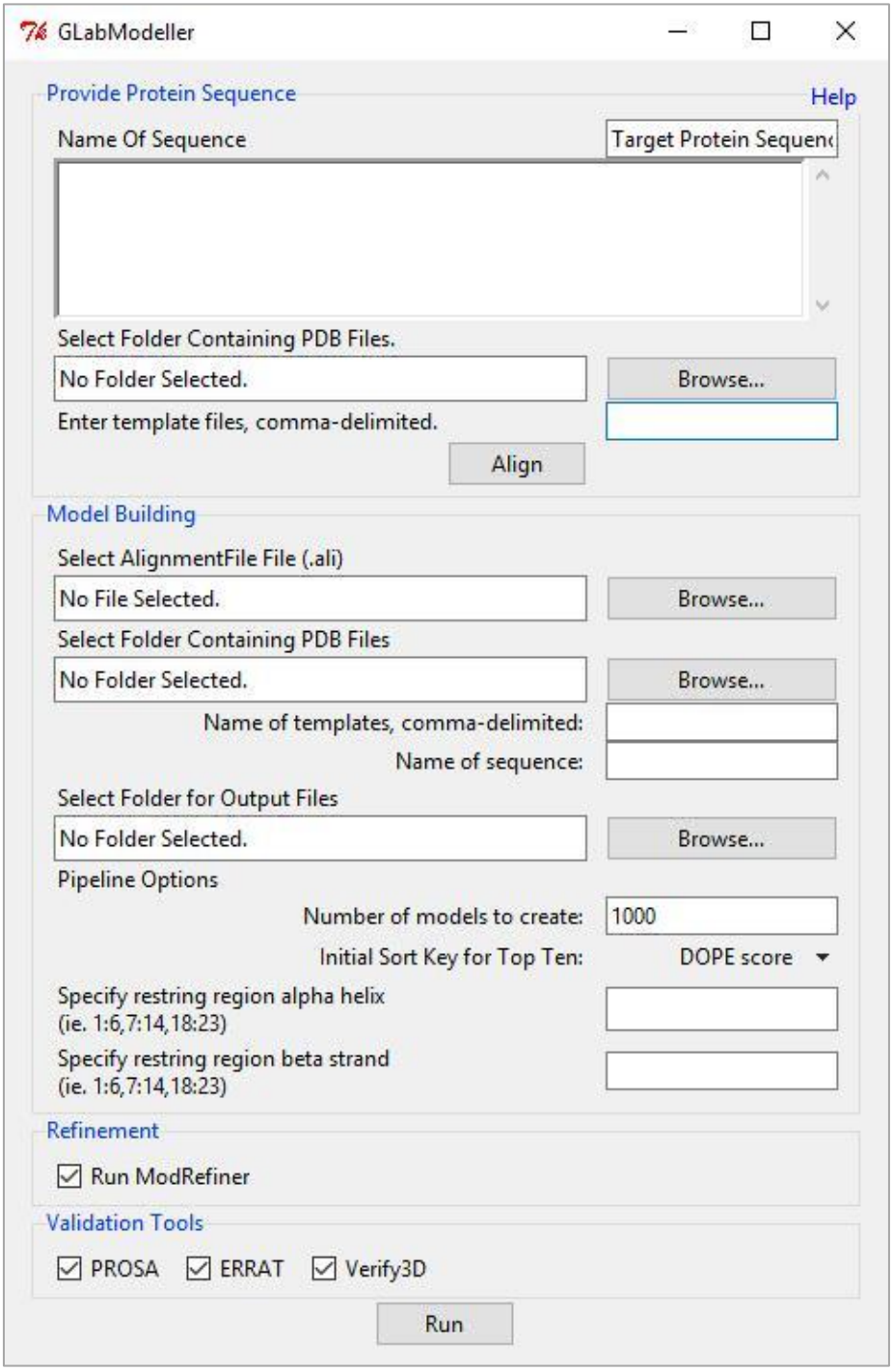


Figure 6.2. Graphical user interference (GUI) for GlabModeller to prepare and run homology modelling.

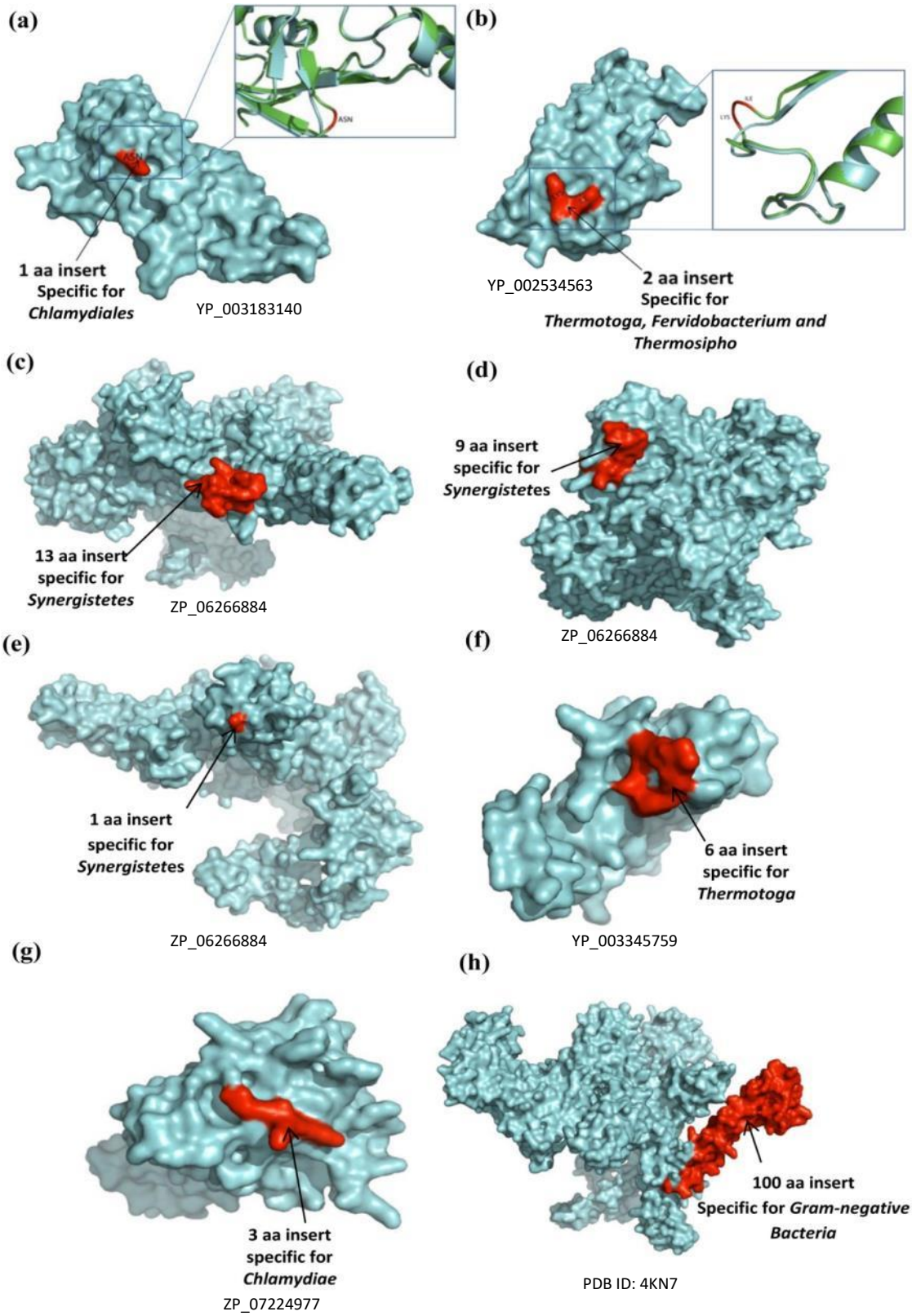

(a)
1 aa insert
Specific for
Chlamydiales
YP_003183140
(b)
YP_002534563
2 aa insert
Specific for
Thermotoga, Fervidobacterium and
Thermosipho
(c)
13 aa insert
specific for
Synergistetes
ZP_06266884
(d)
9 aa insert
specific for
Synergistetes
ZP_06266884
(e)
1 aa insert
specific for
Synergistetes
ZP_06266884
(f)
6 aa insert
specific for
Thermotoga
YP_003345759
(g)
3 aa insert
specific for
Chlamydiae
ZP_07224977
(h)
100 aa insert
Specific for Gram-negative
Bacteria
PDB ID: 4KN7

Figure 6.3. Surface representation of homology models of DNA-dependent RNA polymerase subunit Alpha (RpoA) and DNA-dependent RNA polymerase subunit Beta (RpoB). Homology model of **(a)** *Eggerthella lenta* RpoA (shown in Cyan) with 1 aa insert and **(b)** *Thermotoga neapolitana* RpoA with 2 aa insert, located on the surface exposed loop region (shown in red). The zoom out figures show cartoon representations of the model (Cyan) superimposed to the template (4KN7_A shown as green). **(c).** Homology model of *Thermotoga naphthophila* RpoB with 6 aa insert. **(d).** *Pyramidobacter piscolens* RpoB with 9 aa insert. **(e).** *Pyramidobacter piscolens* RpoB with 1 aa insert. **(f).** *Thermotoga naphthophila* with 6 aa insert. **(g).** *Chlamydia muridara* RpoB with 3 aa insert. **(h).** *Escherichia coli* RpoB with 100 aa insert specific for Proteobacteria, Chlamydiae, CFBG group and Aquificales homologs, for which crystal structural information was available is indicated by PDB ID. The conserved insertions which are located on the surface exposed loop region are shown as red surface. Each model is indicated by accession numbers (see table 2 for additional information) and represented in different orientations in order to provide the best view for the surface location of different CSIs. The figures were generated using PyMOL (The PyMOL Molecular Graphics System, Version 1.5.0.4 Schrödinger, LLC).

Table 1: List of different CSIs specific for the different microbial groups identified in RpoA and RpoB proteins.

	Accession Number	Indel Size	Indels Specificity	Template/ PDB ID:	References (PubMed :ID)
RpoA	WP_013182538	15-17 aa inserts	*Chlamydiales*	4KN7 (chain A)	16079343
	YP_002534563	2 aa inserts	*Thermotoga*, *Fervidobacterium* and *Thermosipho* genera	4KN7_A	24166034
	YP_003183140	1 aa inserts	Clade II Coriobacteriia (*Eggerthella*, *Cryptobacterium*, *Slackia*, and *Gordonibacter*).	4KN7_A	23524353
RpoB	YP_003345759	6 aa ins	*Thermotoga* genus	4KN7 (chain C)	21503713
	ZP_07224977	3 aa ins	*Chlamydiae*	4KN7_C	23060863
	YP_427779	25 aa ins	*Rhodospirillales* species	4KN7_C	18045498
	ZP_01081213	3 aa ins	Clade-C Cyanobacteria	4KN7_C	19622649
	YP_002488751	2 aa ins	Clade I *Micrococcales* species	1YNJ	22390973
	CAA23625	100 aa ins	Gram-negative Bacteria	4KN7_C/1YNJ_C	15179606
	CAA23625	4 aa del	Gram-negative Bacteria	4KN7_C/1YNJ_C	19196760
	ZP_06266884	1 aa ins	*Synergistetes*	4KN7_C	22711299
	ZP_06266884	9 aa ins	*Synergistetes*	4KN7_C	22711299
	ZP_06266884	13 aa ins	*Synergistetes*	4KN7_C	22711299
	ZP_06266884	1 aa del	*Synergistetes*	4KN7_C	22711299

Table 2: Summary of validation results for various CSI-containing RpoA and RpoB homology models.

	Accession No.	Indel Size	Organism	Template -target Sequence identity	RAMPAGE Analysis (%)			ERRAT	VEROFY 3D (%)
					F	A	O		
RpoA	WP_013182538	15-17 aa inserts	*Waddlia chondrophila*	38%	98.3	1.3	0.4	80.090	85.84
	YP_002534563	2 aa inserts	*Thermotoga neapolitana*	41.29%	99.1	0.9	0.0	73.077	89.86
	YP_003183140	1 aa inserts	*Eggerthella lenta*	45%	98.6	1.4	0.0	72.464	92.59
	Template (4KN7_A)	N/D	*Escherichia coli*	N/D	83.8	12.8	3.4	83.828	83.64
RpoB	YP_003345759	6 aa ins	*Thermotoga naphthophila*	95.4%	95.4	3.1	1.5	83.200	70.15
	ZP_07224977	3 aa ins	*Chlamydia muridara*	97.2%	97.2	2.8	0.0	87.20	84.23
	YP_427779	25 aa ins	*Rhodospirillum rubrum*	97.4%	97.4	2.6	0.0	88.722	87.68
	ZP_01081213	3 aa ins	*Synechococcus sp. RS9917*	93.1%	93.1	4.4	2.3	77.629	82.16
	YP_002488751	2 aa ins	*Arthrobacter chlorophenolics*	96.1%	96.1	3.3	0.7	88.915	57.27
	CAA23625**	100 aa ins	*E. coli*	82.4%	82.4	14.2	3.5	72.923	N/A
	CAA23625**	4 aa del	+	+	+	+	+	+	+
	ZP_06266884***	1 aa ins	*Pyramidobacter piscolens*	92.5%	92.5	5.7	1.9	65.735	N/A
	ZP_06266884***	9 aa ins	+	+	+	+	+	+	+
	ZP_06266884***	13 aa ins	+	+	+	+	+	+	+
	ZP_06266884***	1 aa del	+	+	+	+	+	+	+
	Template 1 (4KN7_C)	N/D	*E. coli*	N/D	82.4	14.2	3.5	72.923	N/A
	Template 2 (1YNJ_C)	N/D	*Thermus aquaticus*	N/D	80.5	13.1	6.4	76.155	73.72 %

Note:

F = Favored residues, **A** = Allowed residues, **O** = Outliner residues

N/D (Not defined)

N/A (Not available)

** & *** indicates the same homology model was utilized to represent different CSIs

'+' represents value similar to the previous value.

Table 3: List of my other published articles in which the *GlabModeller* was utilized to generate the homolog models and analyses of the structural location of CSIs in protein structure.

S.No.	References
1	**Khadka, B.,** Chatterjee, T., Gupta, B.P., and Gupta, R.S. (2019). Genomic Analyses Identify Novel Molecular Signatures Specific for the *Caenorhabditis* and other Nematode Taxa Providing Novel Means for Genetic and Biochemical Studies. *Genes* 10, 739.
2	Alnajar, S., **Khadka, B**., and Gupta, R.S. (2017). Ribonucleotide Reductases from Bifidobacteria Contain Multiple Conserved Indels Distinguishing Them from All Other Organisms: *In Silico* Analysis of the Possible Role of a 43 aa Bifidobacteria-Specific Insert in the Class III RNR Homolog. *Front Microbiol*. 8, 1409.
3	**Khadka, B**., Adeolu, M., Blankenship, R.E., and Gupta, R.S. (2017). Novel insights into the origin and diversification of photosynthesis based on analyses of conserved indels in the core reaction center proteins. *Photosynth. Res.* 131, 159-171.
4	Ho, J., Adeolu, M., **Khadka, B**., and Gupta, R.S. (2016). Identification of distinctive molecular traits that are characteristic of the phylum "Deinococcus-Thermus" and distinguish its main constituent groups. *Syst. Appl. Microbiol.* 39, 453-463.
5	Zhang, G., Gao, B., Adeolu, M., **Khadka, B**., and Gupta, R.S. (2016). Phylogenomic Analyses and Comparative Studies on Genomes of the *Bifidobacteriales*: Identification of Molecular Signatures Specific for the Order *Bifidobacteriales* and Its Different Subclades. *Front Microbiol.* 7, 978.
6	Naushad, S., Adeolu, M., Goel, N., **Khadka, B**., Al Dahwi, A., and Gupta, R.S. (2015). Phylogenomic and molecular demarcation of the core members of the polyphyletic *pasteurellaceae* genera *actinobacillus*, *haemophilus*, and *pasteurella*. *Int. J Genomics.* 2015, 198560.
7	Gupta, R.S. and **Khadka, B**. (2015). Evidence for the presence of key chlorophyll-biosynthesis-related proteins in the genus *Rubrobacter* (Phylum Actinobacteria) and its implications for the evolution and origin of photosynthesis. *Photosynth. Res.* 127, 201-218.
8	Hassan, F.M.N. and Gupta, R.S. (2018). Novel Sequence Features of DNA Repair Genes/Proteins from *Deinococcus* Species Implicated in Protection from Oxidatively Generated Damage. *Genes.* 9, 149.
9	Sharma, R. and Gupta, R.S. (2019). Novel Molecular Synapomorphies Demarcate Different Main Groups/Subgroups of *Plasmodium* and Piroplasmida Species Clarifying Their Evolutionary Relationships. *Genes.* 10, 490.

CHAPTER 7

CONCLUSIONS AND FUTURE DIRECTIONS

Research Summary

The exponential growth of genome sequence information from diverse organisms has allowed efficient use of the comparative genomic approaches for the discovery of novel and reliable shared-derived molecular characteristics that serve to clearly demarcate different groups of organisms. One such class of molecular markers that provide powerful means for distinguishing different groups of organisms and to elucidate several of the evolutionary and phylogenetic relationships between them is referred to as Conserved Signature Insertions and Deletions (CSIs) (Rivera and Lake, 1992; Baldauf and Palmer, 1993; Gupta, 1998; Rokas and Holland, 2000; Gupta, 2014). CSIs are insertions or deletions of defined lengths that are present at specific locations within highly conserved regions of widely distributed proteins (Gupta, 1998). CSIs have been identified at multiple phylogenetic depths ranging from those shared by multiple phyla (Griffiths and Gupta, 2007) to those specific to individual strains (Ahmod et al., 2011; Wong et al., 2014). Additionally, CSIs have been used in the past to resolve a number of important evolutionary questions, dating back billions of years. These include clarifying relationships among animals, plants and fungi (Baldauf and Palmer, 1993; Gupta, 1995), identifying closest relatives of the vertebrates as well as the primates (Janecka et al., 2007; Gupta, 2016c), shedding light into the origin and evolution of photosynthetic reaction center core proteins (Gupta, 2012; Gupta and Khadka, 2015; Khadka et al., 2017) and providing insight into the branching order of the main bacterial phyla (Gupta, 2001; Gupta, 2014). In addition, earlier work has shown that CSIs, including the CSIs that have size of 1-2 amino acid in length, in several important proteins (e.g., GroEL, DnaK, GyrB,

etc.) play important functional roles in the CSI-containing organisms, and abolition or any substantial changes in the sequences of CSIs adversely impact cell growth or other critical functions (Chatterji et al., 2000; Singh and Gupta, 2009; Schoeffler et al., 2010; Clarke and Irvine, 2013; Alnajar et al., 2017; Gupta et al., 2017). Moreover, analyses of a number CSIs has shown that the CSIs are generally located within surface loops of the proteins with their residues exposed toward the surface (Hsing and Cherkasov, 2008; Singh and Gupta, 2009; Gupta and Khadka, 2015) and thus are predicted to play important roles in protein-protein and protein-ligand interactions (Akiva et al., 2008; Hashimoto and Panchenko, 2010). In view of the high degree of specificity (i.e. evolutionary conservation) of CSIs for particular group of organisms and their localization in protein structures in surface exposed loops, it is hypothesized that CSIs are involved in ancillary functions that are essential for the CSI-containing group of organisms. However, despite the important predicted role played by the CSIs, the underlying functional mechanism of most of the identified CSIs in different essential genes/proteins remains largely unknown. An understanding of these new ancillary functions as a result of these rare genetic changes in the sequence and structural change is of much importance. This is particularly relevant in light of the worldwide effort by structural genomics which aims to provide a structural representative for most homologous protein families (Baker and Sali, 2001; Chance et al., 2004; Khafizov et al., 2014). A knowledge of three-dimensional structural information of proteins often provides useful insights required to understand the macromolecular function (Terwilliger, 2011; Fajardo and Fiser, 2013) and the analyses of the structural features of CSIs is thus

critical to shed lights into the previously undetected functional relationships of CSIs hidden at the sequence level.

The major focus of my thesis is on using comparative genomic approaches to identify large numbers of CSIs in important proteins which are distinctive characteristics of fungi as well as other important groups of organisms. I have utilized the identified CSIs to understand both the evolutionary history of protein families for delineating relationships between organisms and to investigate the structural and functional aspects of CSIs using various bioinformatics and computational approaches. The evolutionary aspects of my research include work that I have described in Chapter 2 of this thesis. Here, I describe the use of a CSI-based approach in conjunction with species distribution and phylogenetic analysis to provide novel insight into the origin/distribution and evolutionary relationships of the PIP4K/PIP5K protein family.

The members of PIP4K/PIP5K family constitute crucial players in the regulation of phosphatidylinositides, which reside at the core of the phosphatidylinositol signalling pathway, controlling a wide range of fundamental cellular processes in eukaryotes (Martin, 1998; Oude Weernink et al., 2000; van den Bout and Divecha, 2009; Bulley et al., 2015). Despite the important role played by these proteins, there is a limited understanding of the overall evolutionary relationships between different members of PIP4K/PIP5K families and subfamilies of proteins. Additionally, no molecular or biochemical characteristics are known that clearly distinguish the different members of this protein family. In this work, we describe the detailed analysis of the species distribution pattern of PIP4K/PIP5K isozymes/homologs in various eukaryotic lineages

and phylogenetic analysis based on the sequences of these proteins. In parallel, our comparative analysis of the PIP4Ks and PIP5Ks protein sequences have identified six highly-specific molecular markers consisting of CSIs that are uniquely shared by either PIP4K or PIP5K proteins or both, or specific to subfamilies of these proteins. On the basis of the analysis of distribution pattern of these identified CSIs, we were able to reliably determine the specific stages in the evolutionary history of eukaryotic organism, where the gene duplication events leading to the diversification of the PIP4K/PIP5K families and subfamilies of proteins have occurred. In addition, we also explored the structural features and location of all the identified CSIs in the PIP4K/PIP5K family of proteins and showed that all identified CSIs are located on surface exposed loop regions and are thus predicted to perform important roles in mediating novel functional interactions (Khadka and Gupta, 2019).

In the subsequent work, described in Chapter 3 of this thesis, to investigate the structural and functional aspects of CSIs, I analyzed one of the CSIs that we identified in a member of the PIP4K/PIP5K family of proteins. This includes my work on the enzyme PIP5K which play a pivotal role in generating phosphatidylinositol (4,5)-bisphosphate [Ptdlns(4,5)-P2], a key regulator of phosphoinositide signalling pathway (Majerus, 1992; Loijens et al., 1996; van den Bout and Divecha, 2009; Kutateladze, 2010). In yeast, such as *Saccharomyces cerevisiae*, it is the sole PIP5K that controls a wide range of essential cellular functions (Desrivieres et al., 1998; Homma et al., 1998; van den Bout and Divecha, 2009; Guillas et al., 2013). Using a comparative genomic approach we compared the available protein sequences of PIP5Ks form different organisms. This led to

the identification of an 8 aa conserved signature insert in PIP5K protein that is uniquely shared by different species of *Saccharomycetaceae*, but absent in any mammalian or animal homologs. Inserts of smaller sizes are also found in the same position in some other fungi, which are likely specific for other groups or families of fungi. Because the conserved insertions or deletions in protein sequences are predicted to be functionally important for the group of organisms where they are found, I explored the structural features of the identified 8 aa CSI to gain potential insight into its functional significance. As no solved structure of PIP5K was available from fungi, I utilized the homology modelling technique to generate the structural models of PIP5K from *Saccharomyces cerevisiae* by using the available PIP4K structures as a template. Analysis of PIP5K structural model reveals that the *Saccharomycetaceae*-specific 8 aa CSI forms a surface exposed loop region at the surface of this protein, which is present at the same face as the activation loop region of this protein. Furthermore, to investigate its possible role in membrane binding we calculated the electrostatic potential surface (EPS) of PIP5K from *S. cerevisiae.* The EPS analysis shows that the residues from the 8 aa insert contribute toward the formation of a highly positively charged patch on the surface of this protein, through which the electrostatic interaction with anionic head groups of the bilayer membrane, is expected to play a role in the membrane binding. To investigate this prediction, we then utilize the molecular dynamics (MD) simulations to examine the binding interaction of the PIP5K, by creating structural models of PIP5K lacking and containing the conserved insert, using two different membrane lipid bilayer models. The results obtained from the MD simulation provided insights into the underlying

mechanism of interaction of PIP5K to the membrane lipid bilayers and underpinned the idea that the identified fungal-specific 8 aa conserved insert plays an important role in the membrane binding (Khadka and Gupta, 2017).

In addition to the work that has been described in Chapter 3 of this thesis, I have also carried out similar structural and functional studies on a number of CSIs in key proteins specific for two important groups of bacteria that are described in chapter 4 and chapter 5. The work described in Chapter 4 of this thesis is a published work (Gupta et al., 2017), whereas, the work described in Chapter 5 of this thesis has been submitted for publication. A brief summary of these two studies is provided below.

Bifidobacteria comprise an important group of commensal bacteria that forms a significant constituent in the microbiota of humans and other mammals (Biavati et al., 2000; Turroni et al., 2011; Ventura et al., 2014). These bacteria are known for several health-promoting benefits on their hosts (Pokusaeva et al., 2011). Bifidobacteria possesses a unique fermentation pathway known as the “bifid shunt” for the metabolism of different carbohydrates (Meile et al., 2001; Takahashi et al., 2010) which is based on a key enzyme called phosphoketolase (PKs). Unlike phosphoketolase (XPKs) from other bacteria, which shows specificity only for only Xylulose-5-phosphate (X5P), the bifidobacteria phosphoketolase (XFPK) possess an unique ability to metabolize both X5P and Fructose-6-phosphate (F6P) (Meile et al., 2001; Yin et al., 2005; Takahashi et al., 2010; Henard et al., 2015). In this study, we focused on analyzing the sequence features of the phosphoketolases to identify any characteristics that could prove helpful in understanding the differences between the two forms of PKs found in different

organisms. This lead to the identification of multiple highly specific molecular differences in the forms of CSIs that clearly distinguish the phosphoketolase of bifidobacteria from the phosphoketolase homologs found in most other bacteria. Interestingly, we noted that the most of the molecular signatures that are specific for the XFPKs from bifidobacteria were also shared by the PKs from the *Coriobacteriales* order of bacteria, which is comprised of saccharolytic organisms also belonging to the phylum Actinobacteria. Analysis of the branching pattern from phylogenetic tree provided evidence that the PKs in bifidobacteria are specifically related to those found in the *Coriobacteriales*, indicating that the gene for PK (XFPK) was horizontally transferred between these two groups. Additionally, homology modelling, structural analyses, and protein-protein docking studies revealed that the CSIs that are distinguishing features of *Bifidobacteriales*/*Coriobacteriales* are located on the surface exposed loop region at the subunit interface in the XFPK structure and they are indicated to be involved in the formation/stabilization of XFPK dimer.

In another study, we have examined the evolutionary and functional significance of several CSIs in the SecA protein. The SecA, a conserved ATPase, is a key component of bacterial Sec-translocation system (Mori and Ito, 2001; Vrontou and Economou, 2004) that plays a key role in the secretion of a wide variety of bacterial proteins (Schmidt and Kiser, 1999; Gil et al., 2004). Previous work from our lab has identified a large number of CSIs that constituted distinguished molecular characteristics of the order of bacteria which contains some of the most thermophilic and hyperthermophilic species of bacteria known (Griffiths and Gupta, 2004a; Griffiths and Gupta, 2006; Gupta and Bhandari,

2011; Gupta and Lali, 2013). Of the identified CSIs, one CSI constitutes a large 50 aa insert present exclusively in SecA homologs from the orders *Thermotogales* and *Acquificales*. Another CSI comprised of a 76 aa insert in the homologs of SecA is uniquely shared by members of the order *Thermales* and *Hydrogenibacillus schlegelli*. However, it remains unclear how the shared presence of these rare genetic changes in SecA homologs, that are distinctive characteristics of a thermophilic/hyperthermophilic group of bacteria, contribute toward the function of this protein. To investigate this, I carried out the computational analysis of the sequence and structural features of these large CSIs in SecA protein. A phylogenetic tree based on the proteins sequences of SecA homologs was created and analyses of its branching pattern provided evidence that these large CSIs have originated independently in these unrelated phyla of hyperthermophilic bacteria due to their selective advantageous functional roles. As the main commonly shared characteristic of these phyla is their ability to grow at high temperature, it strongly suggests that the presence of these large CSIs in the SecA protein is advantageous or necessary for its functioning at high temperature. Analyses of the amino acid composition and structural features of these CSIs show that these large CSIs constituting mostly of conserved charged amino acid residues are located on the surface exposed region of the SecA. To further investigate the functional significance of these CSIs that are specific for thermophilic/hyperthermophilic bacteria, a comparison of the crystal structure and homology model of *Thermotoga maritima* SecA structure with and without 50 aa CSI was carried out. These comparisons showed that the CSI in the protein was interacting with a number of water molecules and they formed an intermediate interaction between

the insert and adenosine group of ADP molecule. In view of this observation, to investigate the structural dynamics of 50 aa CSI and water molecules, we explore the MD simulation using *Tm*SecA crystal structure at two different temperature settings, (303.15K and 363.15K). The results from MD studies identified a conserved network of stable water molecules near the 50 aa insert. Conserved residues such as GLU185 which is present within the loop region of CSI are found to make a key contribution toward these interactions. The results from this analysis have provided novel insight into the possible role of this CSI and open up an area for future biochemical experiments to be conducted to further clarify the function of these CSIs towards the thermostability of the protein.

Future Directions

The research that I have presented in this thesis provides several intriguing directions for future study. One interesting line of research is to utilize the tools and methods (e.g. the computational pipeline "*GlabModeller*"), for the rapid and efficient mapping of structural location and features of a large number of previously identified or newly discovered CSIs in different proteins. The work from our lab over the past few decades has identified >1000 various CSIs in important proteins which are specifically found in a different group of organism (pathogenic and non-pathogenic). Most of these CSIs are found in highly conserved proteins which are involved in essential functions (Chan et al., 2007). A few examples of such important target proteins where such analysis can be performed include, (i) Recombinase A (RecA), which is essential for the repair and maintenance of DNA (Cox, 2007) (ii) Serine hydroxymethyl transferase

(SHMT), which links amino acid and nucleotide metabolisms by generating key intermediate for one-carbon transfer reactions (Rao et al., 2000), and (iii) Lon Protease, an ATP-dependent protease essential for regulation and energy-dependent degradation of short-lived proteins (Tsilibaris et al., 2006). The primary biochemical functions of these conserved proteins are critical for cell viability and growth and thus are expected to remain the same in all organisms. Therefore a key question that emerges is "*What functional constraints help create and maintain these rare genetic changes in these conserved protein sequence/structure in particular group of organism*?" However, use of experimental approaches to study all the CSIs is a prohibitively time and resource intensive process and it will require extensive work before the functional significance of any of these CSIs is understood. In contrast, the use of computational and bioinformatics approaches allows the rapid analysis of large number of CSIs. For instance, the information obtained by mapping and analysing the structural features of various CSIs can be utilized to unravel their novel roles in protein function by employing additional bioinformatic approaches such as molecular dynamics simulation, genomic organizations of the genes containing these CSI and protein-protein interaction network analysis. Biochemical experiments can then be conducted to further complement the results. In addition to the GlabModeller, our lab has developed a number of computational tools that are useful in this regard. These include tools for the identification of CSIs available on the www.gleans.net website (Gupta, 2014) and tools such as GLIMPSE, a phylogenomic analysis pipeline tool developed by our lab to create supermatrix based large scale phylogenetic trees and to calculate genomic distance using multiple methodologies

(Adeolu et al., 2016). The availability of these resources will pave the path for other researchers to endeavor genome sequence-based evolutionary research, and to search for novel and informative molecular signatures.

Another interesting line of research is to utilize the structural and functional information about CSIs, that I have described in this thesis, to develop a novel and important class of drug targets that can be effectively inhibited by small molecule drug leads (Gupta, 2018). This is particularly relevant as there is a dire need for the development of new antibacterials with a novel mode of action due to the growing health and economic burden caused by the rapid spread in antibiotic resistance bacteria (Payne et al., 2007; Walsh and Wencewicz, 2014; Wright, 2015). Although, the advent of genomic sequencing held the promise of numerous novel and easily identifiable antimicrobial drug targets which could be derived from the comparative analysis of genomic sequence data (Tang and Moxon, 2001). However, there has been limited success in identifying novel drugs using the information based on genomic sequence data (Betz et al., 2005; Shendure et al., 2019). The concept of using CSIs as a drug target stems from the fact that CSIs have been experimentally shown to be essential for the organisms in which they are found and their removal or significant alteration in their sequences are incompatible with the viability of CSI-containing organisms (Singh and Gupta, 2009). Additional features of the CSIs which make them attractive drug target is that most of the CSIs identified are present in conserved proteins that play critical roles in different cellular processes associated with a variety of pathological conditions e.g. PIP5K, RpoB, etc. and which are proven drug targets (Chan et al., 2007; Drake and

Huang, 2014; Gupta, 2018). Further, as the CSIs in proteins are generally found on the surface exposed loops in protein structure which are predicted to play important roles in facilitating novel protein-protein or protein-ligand interactions that are specific for the CSI-containing organisms (Akiva et al., 2008; Hormozdiari et al., 2009; Hashimoto and Panchenko, 2010; Clarke and Irvine, 2013; Khadka and Gupta, 2017; Alnajar et al., 2017; Gupta et al., 2017; Gupta, 2018). Screening for small molecules which bind to the CSIs and inhibit their function(s) should prove growth-inhibitory for the CSI-containing organisms. The idea of using CSIs as drug-target was tested by Nandan *et al.* to screen for compounds which showed inhibitory effects against Elongation factor-1 alpha (EF-1a) by targeting a 12 amino acid deletion in this protein which was specific for the protozoan parasite *Leishmania donovani* (Cherkasov et al., 2005; Lopez et al., 2007; Nandan et al., 2007). Therefore, the application of structural based-drug design approach such as virtual screening and *in vitro* assay of the compounds that will bind to these CSIs and thereby interfere with their cellular functions could lead to the discovery of novel compounds that specifically inhibit the growth of CSI-containing organisms (Gupta, 2018).

Concluding Remarks

The rapid availability of genome sequence data is providing researchers with a plethora of opportunities to understand the evolutionary relationship of different organisms. This rich resource is also enabling the discovery of reliable novel molecular characteristics that are specifically shared by different groups of organisms. The CSIs in protein sequences represent one such important category of molecular markers whose

discovery has been enabled by genomic sequence data and which are used extensively as reliable taxonomic markers and efficient diagnostic markers**.** Recent structural and functional studies on CSIs from our lab and others using various computational and bioinformatics approaches are providing a novel insights towards the structural location, features and cellular and physiological roles for the large number of identified CSIs in various proteins (Alnajar et al., 2017; Gupta et al., 2017; Khadka and Gupta, 2017; Khadka and Gupta, 2019). Future studies aimed at understanding the functions of these CSIs and their applications for the development of novel diagnostics as well as their potential use as novel drug targets should be of great interest.

BIBLIOGRAPHY

Abrahama, M.J., Schulzb, R., Pálla, S., Smith, J.C., Hessa, B., and Lindahla, E. (2016). GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. SoftwareX *1-2*, 19-25.

Adeolu, M., Alnajar, S., Naushad, S., and Gupta, S. (2016). Genome-based phylogeny and taxonomy of the *'Enterobacteriales*': proposal for *Enterobacterales* ord. nov. divided into the families *Enterobacteriaceae, Erwiniaceae* fam. nov., *Pectobacteriaceae* fam. nov., *Yersiniaceae* fam. nov., *Hafniaceae* fam. nov., *Morganellaceae* fam. nov., and *Budviciaceae* fam. nov. Int. J Syst. Evol Microbiol. *66*, 5575-5599.

Ahmod, N.Z., Gupta, R.S., and Shah, H.N. (2011). Identification of a *Bacillus anthracis* specific indel in the yeaC gene and development of a rapid pyrosequencing assay for distinguishing *B. anthracis* from the B. cereus group. J. Microbiol. Methods *87*, 278-285.

Ajawatanawong, P. and Baldauf, S.L. (2013). Evolution of protein indels in plants, animals and fungi. BMC. Evol Biol *13*, 140.

Akiva, E., Itzhaki, Z., and Margalit, H. (2008). Built-in loops allow versatility in domain-domain interactions: lessons from self-interacting domains. Proc. Natl. Acad. Sci. U. S. A *105*, 13292-13297.

Alm, R.A., Ling, L.S., Moir, D.T., King, B.L., Brown, E.D., Doig, P.C., Smith, D.R., Noonan, B., Guild, B.C., deJonge, B.L., Carmel, G., Tummino, P.J., Caruso, A., Uria-Nickelsen, M., Mills, D.M., Ives, C., Gibson, R., Merberg, D., Mills, S.D., Jiang, Q., Taylor, D.E., Vovis, G.F., and Trust, T.J. (1999). Genomic-sequence comparison of two unrelated isolates of the human gastric pathogen *Helicobacter pylori*. Nature *397*, 176-180.

Almassy, R.J. and Dickerson, R.E. (1978). Pseudomonas cytochrome c551 at 2.0 A resolution: enlargement of the cytochrome c family. Proc Natl. Acad. Sci. U. S. A *75*, 2674-2678.

Alnajar, S., Khadka, B., and Gupta, R.S. (2017). Ribonucleotide Reductases from Bifidobacteria Contain Multiple Conserved Indels Distinguishing Them from All Other Organisms: *In Silico* Analysis of the Possible Role of a 43 aa Bifidobacteria-Specific Insert in the Class III RNR Homolog. Front Microbiol. *8*, 1409.

Altschul, S.F., Gertz, E.M., Agarwala, R., Schaffer, A.A., and Yu, Y.K. (2009). PSI-BLAST pseudocounts and the minimum description length principle. Nucleic Acids Res. *37*, 815-824.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and 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., and Lipman, D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. *25*, 3389-3402.

Anfinsen, C.B. (1959). The Molecular Basis of Evolution. (New York: Wiley).

Anfinsen, C.B. (1973). Principles that govern the folding of protein chains. Science *181*, 223-230.

Anfinsen, C.B. and Haber, E. (1961). Studies on the reduction and re-formation of protein disulfide bonds. J Biol Chem *236*, 1361-1363.

Anfinsen, C.B., Haber, E., Sela, M., and White, F.H.J. (1961). The kinetics of formation of native ribonuclease during oxidation of the reduced polypeptide chain. Natl. Acad. Sci *47*, 1314.

Aspnas, M., Mattila, K., Osowski, K., and Westerholm, J. (2010). Code optimization of the subroutine to remove near identical matches in the sequence database homology search tool PSI-BLAST. J Comput. Biol. *17*, 819-823.

Avey, H.P., Boles, M.O., Carlisle, C.H., Evans, S.A., Morris, S.J., Palmer, R.A., Woolhouse, B.A., and Shall, S. (1967). Structure of ribonuclease. Nature *213*, 557-562.

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

Baldauf, S.L. and Palmer, J.D. (1993). Animals and fungi are each other's closest relatives: congruent evidence from multiple proteins. Proc. Natl. Acad. Sci. U. S. A *90*, 11558-11562.

Balla, T. (2013). Phosphoinositides: tiny lipids with giant impact on cell regulation. Physiol Rev. *93*, 1019-1137.

Balla, T., Szentpetery, Z., and Kim, Y.J. (2009). Phosphoinositide signaling: new tools and insights. Physiology. (Bethesda.) *24*, 231-244.

Belda, E., Moya, A., and Silva, F.J. (2005). Genome rearrangement distances and gene order phylogeny in gamma-Proteobacteria. Mol. Biol. Evol *22*, 1456-1467.

Benner, S.A., Cohen, M.A., and Gonnet, G.H. (1993). Empirical and structural models for insertions and deletions in the divergent evolution of proteins. J Mol. Biol. *229*, 1065-1082.

Berman, H., Henrick, K., and Nakamura, H. (2003). Announcing the worldwide Protein Data Bank. Nat. Struct. Biol *10*, 980.

Bermudez, M., Mortier, J., Rakers, C., Sydow, D., and Wolber, G. (2016). More than a look into a crystal ball: protein structure elucidation guided by molecular dynamics simulations. Drug Discov. Today *21*, 1799-1805.

Bernal, J.D. and Crowfoot, D. (1934). X-ray Photographs of Crystalline Pepsin. Nature *133*, 794-795.

Bernard, B.R., Bruccoleri, R.E., Olafson, B.D., States, D.J., Swaminathan, S., and Karplus, M. (1983). CHARMM: A program for macromolecular energy, minimization, and dynamics calculations. J. Comput. Chem *4*, 187-217.

Bernstein, F.C., Koetzle, T.F., Williams, G.J., Meyer, E.F., Jr., Brice, M.D., Rodgers, J.R., Kennard, O., Shimanouchi, T., and Tasumi, M. (1977). The Protein Data Bank: a computer-based archival file for macromolecular structures. J Mol. Biol. *112*, 535-542.

Betz, U.A., Farquhar, R., and Ziegelbauer, K. (2005). Genomics: success or failure to deliver drug targets? Curr. Opin. Chem Biol. *9*, 387-391.

Biavati, B., Vescovo, M., Torriani, S., and Bottazzi, V. (2000). Bifidobacteria: history, ecology, physiology and applications. Annals of Microbiology *50*, 117-131.

Bielawski, J.P. and Yang, Z. (2004). A maximum likelihood method for detecting functional divergence at individual codon sites, with application to gene family evolution. J Mol. Evol *59*, 121-132.

Biggin, P.C. and Bond, P.J. (2015). Molecular dynamics simulations of membrane proteins. Methods Mol. Biol. *1215*, 91-108.

Bilofsky, H.S. and Burks, C. (1988). The GenBank genetic sequence data bank. Nucleic Acids Res. *16*, 1861-1863.

Blake, C.C., Koenig, D.F., Mair, G.A., North, A.C., Phillips, D.C., and Sarma, V.R. (1965). Structure of hen egg-white lysozyme. A three-dimensional Fourier synthesis at 2 Angstrom resolution. Nature *206*, 757-761.

Boehr, D.D., Nussinov, R., and Wright, P.E. (2009). The role of dynamic conformational ensembles in biomolecular recognition. Nat. Chem. Biol. *5*, 789-796.

Bohm, H.J. (1994). The development of a simple empirical scoring function to estimate the binding constant for a protein-ligand complex of known three-dimensional structure. J Comput. Aided Mol. Des *8*, 243-256.

Bowie, J.U., Luthy, R., and Eisenberg, D. (1991). A method to identify protein sequences that fold into a known three-dimensional structure. Science *253*, 164-170.

Bragg, L., Kendrew J.C, and Perutz M.F (1950). Polypeptide chain configurations in crystalline proteins. Proc. R. Soc. Lond. Ser. A Math. Phys. Sci. *203*, 321-357.

Bragg, W.H. and Bragg, W.L. (1913). The reflection of X-rays by crystals. Proc R Soc *A88*, 428-438.

Brown, H., Sanger, F., and Kitai, R. (1955). The structure of pig and sheep insulins. Biochem J. *60*, 556-565.

Brown, J.R. and Auger, K.R. (2011). Phylogenomics of phosphoinositide lipid kinases: perspectives on the evolution of second messenger signaling and drug discovery. BMC. Evol Biol. *11*, 4.

Browne, W.J., North, A.C., Phillips, D.C., Brew, K., Vanaman, T.C., and Hill, R.L. (1969). A possible three-dimensional structure of bovine alpha-lactalbumin based on that of hen's egg-white lysozyme. J. Mol. Biol. *42*, 65-86.

Buch, I., Giorgino, T., and De Fabritiis, G. (2011). Complete reconstruction of an enzyme-inhibitor binding process by molecular dynamics simulations. Proc. Natl. Acad. Sci. U. S. A *108*, 10184-10189.

Bulley, S.J., Clarke, J.H., Droubi, A., Giudici, M.L., and Irvine, R.F. (2015). Exploring phosphatidylinositol 5-phosphate 4-kinase function. Adv. Biol. Regul. *57*, 193-202.

Burks, C., Fickett, J.W., Goad, W.B., Kanehisa, M., Lewitter, F.I., Rindone, W.P., Swindell, C.D., Tung, C.S., and Bilofsky, H.S. (1985). The GenBank nucleic acid sequence database. Comput. Appl. Biosci. *1*, 225-233.

Burley, S.K., Berman, H.M., Bhikadiya, C., Bi, C., Chen, L., Di Costanzo, L., Christie, C., Dalenberg, K., Duarte, J.M., Dutta, S., Feng, Z., Ghosh, S., Goodsell, D.S., Green, R.K., Guranovic, V., Guzenko, D., Hudson, B.P., Kalro, T., Liang, Y., Lowe, R., Namkoong, H., Peisach, E., Periskova, I., Prlic, A., Randle, C., Rose, A., Rose, P., Sala, R., Sekharan, M., Shao, C., Tan, L., Tao, Y.P., Valasatava, Y., Voigt, M., Westbrook, J., Woo, J., Yang, H., Young, J., Zhuravleva, M., and Zardecki, C. (2019). RCSB Protein Data Bank: biological macromolecular structures enabling research and education in fundamental biology, biomedicine, biotechnology and energy. Nucleic Acids Res. *47*, D464-D474.

Butler, B. (1993). An overview of computer software developed to search biological sequence databases. Antisense Res. Dev. *3*, 243-252.

Cameron, G.N. (1988). The EMBL data library. Nucleic Acids Res. *16*, 1865-1867.

Campbell, A., Mrazek, J., and Karlin, S. (1999). Genome signature comparisons among prokaryote, plasmid, and mitochondrial DNA. Proc. Natl. Acad. Sci. U. S. A *96*, 9184-9189.

Campbell, E.A., Korzheva, N., Mustaev, A., Murakami, K., Nair, S., Goldfarb, A., and Darst, S.A. (2001). Structural mechanism for rifampicin inhibition of bacterial rna polymerase. Cell *104*, 901-912.

Campbell, I.D. (2002). Timeline: the march of structural biology. Nat. Rev. Mol. Cell Biol *3*, 377-381.

Cantarel, B.L., Morrison, H.G., and Pearson, W. (2006). Exploring the relationship between sequence similarity and accurate phylogenetic trees. Mol. Biol. Evol *23*, 2090-2100.

Capener, C.E., Shrivastava, I.H., Ranatunga, K.M., Forrest, L.R., Smith, G.R., and Sansom, M.S. (2000). Homology modeling and molecular dynamics simulation studies of an inward rectifier potassium channel. Biophys. J. *78*, 2929-2942.

Caporaso, J.G., Lauber, C.L., Walters, W.A., Berg-Lyons, D., Huntley, J., Fierer, N., Owens, S.M., Betley, J., Fraser, L., Bauer, M., Gormley, N., Gilbert, J.A., Smith, G., and Knight, R. (2012). Ultra-high-throughput microbial community analysis on the Illumina HiSeq and MiSeq platforms. ISME. J *6*, 1621-1624.

Carlsson, J., Coleman, R.G., Setola, V., Irwin, J.J., Fan, H., Schlessinger, A., Sali, A., Roth, B.L., and Shoichet, B.K. (2011). Ligand discovery from a dopamine D3 receptor homology model and crystal structure. Nat. Chem. Biol. *7*, 769-778.

Carrascoza, F., Zaric, S., and Silaghi-Dumitrescu, R. (2014). Computational study of protein secondary structure elements: Ramachandran plots revisited. J Mol. Graph. Model. *50*, 125-133.

Case, D.A., Cheatham, T.E., III, Darden, T., Gohlke, H., Luo, R., Merz, K.M., Jr., Onufriev, A., Simmerling, C., Wang, B., and Woods, R.J. (2005). The Amber biomolecular simulation programs. J Comput. Chem *26*, 1668-1688.

Ceulemans, H. and Russell, R.B. (2004). Fast fitting of atomic structures to low-resolution electron density maps by surface overlap maximization. J. Mol. Biol. *338*, 783-793.

Chan, S.K., Hsing, M., Hormozdiari, F., and Cherkasov, A. (2007). Relationship between insertion/deletion (indel) frequency of proteins and essentiality. BMC. Bioinformatics. *8*, 227.

Chance, M.R., Fiser, A., Sali, A., Pieper, U., Eswar, N., Xu, G., Fajardo, J.E., Radhakannan, T., and Marinkovic, N. (2004). High-throughput computational and experimental techniques in structural genomics. Genome Res. *14*, 2145-2154.

Chang, B.S., Ugalde, J.A., and Matz, M.V. (2005). Applications of ancestral protein reconstruction in understanding protein function: GFP-like proteins. Methods Enzymol. *395*, 652-670.

Charlesworth, D., Charlesworth, B., and McVean, G.A. (2001). Genome sequences and evolutionary biology, a two-way interaction. Trends Ecol. Evol *16*, 235-242.

Chatterji, M., Unniraman, S., Maxwell, A., and Nagaraja, V. (2000). The additional 165 amino acids in the B protein of *Escherichia coli* DNA gyrase have an important role in DNA binding. J Biol. Chem *275*, 22888-22894.

Chavent, M., Duncan, A.L., and Sansom, M.S. (2016). Molecular dynamics simulations of membrane proteins and their interactions: from nanoscale to mesoscale. Curr. Opin. Struct. Biol. *40*, 8-16.

Chelliah, V., Chen, L., Blundell, T.L., and Lovell, S.C. (2004). Distinguishing structural and functional restraints in evolution in order to identify interaction sites. J Mol. Biol. *342*, 1487-1504.

Chen, P.E., Geballe, M.T., Stansfeld, P.J., Johnston, A.R., Yuan, H., Jacob, A.L., Snyder, J.P., Traynelis, S.F., and Wyllie, D.J. (2005). Structural features of the glutamate binding site in recombinant NR1/NR2A N-methyl-D-aspartate receptors determined by site-directed mutagenesis and molecular modeling. Mol. Pharmacol. *67*, 1470-1484.

Chen, R. and Weng, Z. (2003). A novel shape complementarity scoring function for protein-protein docking. Proteins *51*, 397-408.

Chen, Y., Bauer, B.W., Rapoport, T.A., and Gumbart, J.C. (2015). Conformational Changes of the Clamp of the Protein Translocation ATPase SecA. J Mol. Biol. *427*, 2348-2359.

Chenna, R., Sugawara, H., Koike, T., Lopez, R., Gibson, T.J., Higgins, D.G., and Thompson, J.D. (2003). Multiple sequence alignment with the Clustal series of programs. Nucleic Acids Res. *31*, 3497-3500.

Cherkasov, A., Nandan, D., and Reiner, N.E. (2005). Selective targeting of indel-inferred differences in spatial structures of highly homologous proteins. Proteins *58*, 950-954.

Chlenov, M., Masuda, S., Murakami, K.S., Nikiforov, V., Darst, S.A., and Mustaev, A. (2005). Structure and function of lineage-specific sequence insertions in the bacterial RNA polymerase beta' subunit. J Mol. Biol. *353*, 138-154.

Chmiel, A.A., Bujnicki, J.M., and Skowronek, K.J. (2005). A homology model of restriction endonuclease SfiI in complex with DNA. BMC. Struct. Biol. *5*, 2.

Chothia, C. and Lesk, A.M. (1986). The relation between the divergence of sequence and structure in proteins. EMBO J. *5*, 823-826.

Chow, C.C., Chow, C., Raghunathan, V., Huppert, T.J., Kimball, E.B., and Cavagnero, S. (2003). Chain length dependence of apomyoglobin folding: structural evolution from misfolded sheets to native helices. Biochemistry *42*, 7090-7099.

Chun, J., Oren, A., Ventosa, A., Christensen, H., Arahal, D.R., da Costa, M.S., Rooney, A.P., Yi, H., Xu, X.W., De Meyer, S., and Trujillo, M.E. (2018). Proposed minimal standards for the use of genome data for the taxonomy of prokaryotes. Int. J Syst. Evol Microbiol. *68*, 461-466.

Chun, J. and Rainey, F.A. (2014). Integrating genomics into the taxonomy and systematics of the Bacteria and Archaea. Int. J Syst. Evol Microbiol. *64*, 316-324.

Clark, M.M., Hildreth, A., Batalov, S., Ding, Y., Chowdhury, S., Watkins, K., Ellsworth, K., Camp, B., Kint, C.I., Yacoubian, C., Farnaes, L., Bainbridge, M.N., Beebe, C., Braun, J.J.A., Bray, M., Carroll, J., Cakici, J.A., Caylor, S.A., Clarke, C., Creed, M.P., Friedman, J., Frith, A., Gain, R., Gaughran, M., George, S., Gilmer, S., Gleeson, J., Gore, J., Grunenwald, H., Hovey, R.L., Janes, M.L., Lin, K., McDonagh, P.D., McBride, K., Mulrooney, P., Nahas, S., Oh, D., Oriol, A., Puckett, L., Rady, Z., Reese, M.G., Ryu, J., Salz, L., Sanford, E., Stewart, L., Sweeney, N., Tokita, M., Van Der, K.L., White, S., Wigby, K., Williams, B., Wong, T., Wright, M.S., Yamada, C., Schols, P., Reynders, J., Hall, K., Dimmock, D., Veeraraghavan, N., Defay, T., and Kingsmore, S.F. (2019). Diagnosis of genetic diseases in seriously ill children by rapid whole-genome sequencing and automated phenotyping and interpretation. Sci. Transl. Med. *11*.

Clarke, J.H. and Irvine, R.F. (2013). Evolutionarily conserved structural changes in phosphatidylinositol 5-phosphate 4-kinase (PI5P4K) isoforms are responsible for differences in enzyme activity and localization. Biochem. J *454*, 49-57.

Cohn, F. (1875). Untersuchungen uber Bakterien. II. Beitr Biol P?anz *3*, 141-208.

Colovos, C. and Yeates, T.O. (1993). Verification of protein structures: patterns of nonbonded atomic interactions. Protein Sci. *2*, 1511-1519.

Comeau, S.R., Gatchell, D.W., Vajda, S., and Camacho, C.J. (2004). ClusPro: an automated docking and discrimination method for the prediction of protein complexes. Bioinformatics *20*, 45-50.

Connell, C., Fung, S., Heiner, C., Bridgham, J., Chakerian, V., Heron, E., Jones, B., Menchen, S., Mordan, W., Raff, M., Recknor, M., Smith, L., Springer, J., Woo, S., and Hunkapiller, M. (1987). Automated DNA sequence analysis. Biotechniques 342-348.

Cox, M.M. (2007). Regulation of bacterial RecA protein function. Crit Rev Biochem Mol. Biol *42*, 41-63.

Crick, F.H. (1958). On protein synthesis. Symp Soc Exp Biol *12*, 138-163.

Darwin, C. (1859). The Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life. (London: John Murray).

Davis, A.M., St Gallay, S.A., and Kleywegt, G.J. (2008). Limitations and lessons in the use of X-ray structural information in drug design. Drug Discov. Today *13*, 831-841.

Dayhoff, M.O. (1965). Computer aids to protein sequence determination. J Theor Biol. *8*, 97-112.

de Graaf, C., Oostenbrink, C., Keizers, P.H., Vugt-Lussenburg, B.M., van Waterschoot, R.A., Tschirret-Guth, R.A., Commandeur, J.N., and Vermeulen, N.P. (2007). Molecular modeling-guided site-directed mutagenesis of cytochrome P450 2D6. Curr. Drug Metab *8*, 59-77.

de Queiroz, A. and Gatesy, J. (2007). The supermatrix approach to systematics. Trends Ecol. Evol *22*, 34-41.

De Rienzo, F., Fanelli, F., Menziani, M.C., and De Benedetti, P.G. (2000). Theoretical investigation of substrate specificity for cytochromes P450 IA2, P450 IID6 and P450 IIIA4. J. Comput. Aided Mol. Des *14*, 93-116.

Desrivieres, S., Cooke, F.T., Parker, P.J., and Hall, M.N. (1998). MSS4, a phosphatidylinositol-4-phosphate 5-kinase required for organization of the actin cytoskeleton in *Saccharomyces cerevisiae*. J. Biol. Chem. *273*, 15787-15793.

Di Paolo, G. and De Camilli, P. (2006). Phosphoinositides in cell regulation and membrane dynamics. Nature *443*, 651-657.

Donoghue, M.J. and Mathews, S. (1998). Duplicate genes and the root of angiosperms, with an example using phytochrome sequences. Mol. Phylogenet. Evol *9*, 489-500.

Doolittle, R.F. and Blombaeck, B. (1964). Amino-Acid Sequence Investigations of Fibrinopeptides from various Mammals: Evolutionary Implications. Nature *202*, 147-152.

Drake, J.M. and Huang, J. (2014). PIP5K1alpha inhibition as a therapeutic strategy for prostate cancer. Proc Natl. Acad. Sci U. S. A *111*, 12578-12579.

Dror, R.O., Arlow, D.H., Maragakis, P., Mildorf, T.J., Pan, A.C., Xu, H., Borhani, D.W., and Shaw, D.E. (2011). Activation mechanism of the beta2-adrenergic receptor. Proc. Natl. Acad. Sci. U. S. A *108*, 18684-18689.

Ebright, R.H. and Busby, S. (1995). The *Escherichia coli* RNA polymerase alpha subunit: structure and function. Curr. Opin. Genet. Dev. *5*, 197-203.

Eck, R.V. (1962). A simplified strategy for sequence analysis of large proteins. Nature *193*, 241-243.

Eck, R.V. and Dayhoff, M.O. (1966). Evolution of the structure of ferredoxin based on living relics of primitive amino Acid sequences. Science *152*, 363-366.

Edgar, R.C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. *32*, 1792-1797.

Edman, P. and Begg, G. (1967). A protein sequenator. Eur. J Biochem *1*, 80-91.

Efron, B. (1992). Bootstrap Methods: Another Look at the Jackknife. In Breakthroughs in Statistics , Samuel Kotz and Norman L.Johnson, eds. (NY: Springer), pp. 569-593.

Eid, J., Fehr, A., Gray, J., Luong, K., Lyle, J., Otto, G., Peluso, P., Rank, D., Baybayan, P., Bettman, B., Bibillo, A., Bjornson, K., Chaudhuri, B., Christians, F., Cicero, R., Clark, S., Dalal, R., Dewinter, A., Dixon, J., Foquet, M., Gaertner, A., Hardenbol, P., Heiner, C., Hester, K., Holden, D., Kearns, G., Kong, X., Kuse, R., Lacroix, Y., Lin, S., Lundquist, P., Ma, C., Marks, P., Maxham, M., Murphy, D., Park, I., Pham, T., Phillips, M., Roy, J., Sebra, R., Shen, G., Sorenson, J., Tomaney, A., Travers, K., Trulson, M., Vieceli, J., Wegener, J., Wu, D., Yang, A., Zaccarin, D., Zhao, P., Zhong, F., Korlach, J., and Turner, S. (2009). Real-time DNA sequencing from single polymerase molecules. Science *323*, 133-138.

Eisenberg, D., Luthy, R., and Bowie, J.U. (1997). VERIFY3D: assessment of protein models with three-dimensional profiles. Methods Enzymol. *277*, 396-404.

Emerling, B.M., Hurov, J.B., Poulogiannis, G., Tsukazawa, K.S., Choo-Wing, R., Wulf, G.M., Bell, E.L., Shim, H.S., Lamia, K.A., Rameh, L.E., Bellinger, G., Sasaki, A.T., Asara, J.M., Yuan, X., Bullock, A., Denicola, G.M., Song, J., Brown, V.,

Signoretti, S., and Cantley, L.C. (2013). Depletion of a putatively druggable class of phosphatidylinositol kinases inhibits growth of p53-null tumors. Cell *155*, 844-857.

Emmert-Streib, F., Dehmer, M., and Yli-Harja, O. (2017). Lessons from the Human Genome Project: Modesty, Honesty, and Realism. Front Genet. *8*, 184.

Epand, R.M. (2017). Features of the Phosphatidylinositol Cycle and its Role in Signal Transduction. J Membr. Biol. *250*, 353-366.

Erdin, S., Ward, R.M., Venner, E., and Lichtarge, O. (2010). Evolutionary trace annotation of protein function in the structural proteome. J Mol. Biol. *396*, 1451-1473.

Fajardo, J.E. and Fiser, A. (2013). Protein structure based prediction of catalytic residues. BMC. Bioinformatics *14*, 63.

Felsenstein, J. (1981). Evolutionary trees from DNA sequences: a maximum likelihood approach. J Mol. Evol *17*, 368-376.

Felsenstein, J. (1985). Confidence Limits on Phylogenies: An Approach Using the Bootstrap. Evolution *39*, 783-791.

Fiser, A. (2004). Protein structure modeling in the proteomics era. Expert. Rev Proteomics. *1*, 97-110.

Fiser, A. (2010). Template-based protein structure modeling. Methods Mol. Biol *673*, 73-94.

Fitch, W.M. (1971). Toward defining the course of evolution: minimum change for a specific tree topology. Systematic Biology *20*, 406-416.

Fitch, W.M. (2000). Homology a personal view on some of the problems. Trends Genet. *16*, 227-231.

Fitch, W.M. and Margoliash, E. (1967). Construction of phylogenetic trees. Science *155*, 279-284.

Fleischmann, R.D., Adams, M.D., White, O., Clayton, R.A., Kirkness, E.F., Kerlavage, A.R., Bult, C.J., Tomb, J.F., Dougherty, B.A., Merrick, J.M., and . (1995). Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. Science *269*, 496-512.

Freddolino, P.L. and Schulten, K. (2009). Common structural transitions in explicit-solvent simulations of villin headpiece folding. Biophys J *97*, 2338-2347.

Friedrichs, M.S., Eastman, P., Vaidyanathan, V., Houston, M., Legrand, S., Beberg, A.L., Ensign, D.L., Bruns, C.M., and Pande, V.S. (2009). Accelerating molecular dynamic simulation on graphics processing units. J Comput. Chem *30*, 864-872.

Friesner, R.A., Banks, J.L., Murphy, R.B., Halgren, T.A., Klicic, J.J., Mainz, D.T., Repasky, M.P., Knoll, E.H., Shelley, M., Perry, J.K., Shaw, D.E., Francis, P., and Shenkin, P.S. (2004). Glide: a new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy. J. Med. Chem. *47*, 1739-1749.

Furnham, N., Dawson, N.L., Rahman, S.A., Thornton, J.M., and Orengo, C.A. (2016). Large-Scale Analysis Exploring Evolution of Catalytic Machineries and Mechanisms in Enzyme Superfamilies. J Mol. Biol. *428*, 253-267.

Furnham, N., Sillitoe, I., Holliday, G.L., Cuff, A.L., Laskowski, R.A., Orengo, C.A., and Thornton, J.M. (2012). Exploring the evolution of novel enzyme functions within structurally defined protein superfamilies. PLoS. Comput. Biol. *8*, e1002403.

Gabaldon, T. and Koonin, E.V. (2013). Functional and evolutionary implications of gene orthology. Nat. Rev. Genet. *14*, 360-366.

Gao, B. and Gupta, R.S. (2012). Phylogenetic framework and molecular signatures for the main clades of the phylum Actinobacteria. Microbiol. Mol. Biol. Rev. *76*, 66-112.

Geszvain, K., Gruber, T.M., Mooney, R.A., Gross, C.A., and Landick, R. (2004). A hydrophobic patch on the flap-tip helix of *E.coli* RNA polymerase mediates sigma(70) region 4 function. J. Mol. Biol. *343*, 569-587.

Gherardini, P.F. and Helmer-Citterich, M. (2008). Structure-based function prediction: approaches and applications. Brief. Funct. Genomic. Proteomic. *7*, 291-302.

Gil, R., Silva, F.J., Pereto, J., and Moya, A. (2004). Determination of the core of a minimal bacterial gene set. Microbiol. Mol. Biol. Rev. *68*, 518-37, table.

Gilbert, W. and Maxam, A. (1973). The nucleotide sequence of the lac operator. Proc. Natl. Acad. Sci. U. S. A *70*, 3581-3584.

Ginalski, K. (2006). Comparative modeling for protein structure prediction. Curr. Opin. Struct. Biol. *16*, 172-177.

Gohlke, H., Hendlich, M., and Klebe, G. (2000). Knowledge-based scoring function to predict protein-ligand interactions. J Mol. Biol. *295*, 337-356.

Goodsell, D.S. and Olson, A.J. (1990). Automated docking of substrates to proteins by simulated annealing. Proteins *8*, 195-202.

Gray, J.J., Moughon, S., Wang, C., Schueler-Furman, O., Kuhlman, B., Rohl, C.A., and Baker, D. (2003). Protein-protein docking with simultaneous optimization of rigid-body displacement and side-chain conformations. J Mol. Biol. *331*, 281-299.

Greer, J. (1981). Comparative model-building of the mammalian serine proteases. J Mol. Biol *153*, 1027-1042.

Griffiths, E. and Gupta, R.S. (2004a). Distinctive protein signatures provide molecular markers and evidence for the monophyletic nature of the deinococcus-thermus phylum. J Bacteriol. *186*, 3097-3107.

Griffiths, E. and Gupta, R.S. (2004b). Signature sequences in diverse proteins provide evidence for the late divergence of the Order *Aquificales*. Int. Microbiol. *7*, 41-52.

Griffiths, E. and Gupta, R.S. (2006). Molecular signatures in protein sequences that are characteristics of the phylum Aquificae. Int. J Syst. Evol Microbiol. *56*, 99-107.

Griffiths, E. and Gupta, R.S. (2007). Phylogeny and shared conserved inserts in proteins provide evidence that *Verrucomicrobia* are the closest known free-living relatives of chlamydiae. Microbiology *153*, 2648-2654.

Gross, C.A., Chan, C.L., and Lonetto, M.A. (1996). A structure/function analysis of *Escherichia coli* RNA polymerase. Philos. Trans. R. Soc. Lond B Biol. Sci. *351*, 475-482.

Gu, Y., Li, D.W., and Bruschweiler, R. (2015). Decoding the Mobility and Time Scales of Protein Loops. J Chem Theory. Comput. *11*, 1308-1314.

Gu, Y., Shrivastava, I.H., Amara, S.G., and Bahar, I. (2009). Molecular simulations elucidate the substrate translocation pathway in a glutamate transporter. Proc. Natl. Acad. Sci. U. S. A *106*, 2589-2594.

Guillas, I., Vernay, A., Vitagliano, J.J., and Arkowitz, R.A. (2013). Phosphatidylinositol 4,5-bisphosphate is required for invasive growth in *Saccharomyces cerevisiae*. J. Cell Sci. *126*, 3602-3614.

Gupta, R.S. (1995). Phylogenetic analysis of the 90 kD heat shock family of protein sequences and an examination of the relationship among animals, plants, and fungi species. Mol. Biol. Evol *12*, 1063-1073.

Gupta, R.S. (1998). Protein phylogenies and signature sequences: A reappraisal of evolutionary relationships among archaebacteria, eubacteria, and eukaryotes. Microbiol. Mol. Biol. Rev. *62*, 1435-1491.

Gupta, R.S. (2001). The branching order and phylogenetic placement of species from completed bacterial genomes, based on conserved indels found in various proteins. Int. Microbiol. *4*, 187-202.

Gupta, R.S. (2004). The phylogeny and signature sequences characteristics of *Fibrobacteres, Chlorobi*, and *Bacteroidetes*. Crit Rev. Microbiol. *30*, 123-143.

Gupta, R.S. (2010). Molecular signatures for the main phyla of photosynthetic bacteria and their subgroups. Photosynth. Res. *104*, 357-372.

Gupta, R.S. (2012). Origin and spread of photosynthesis based upon conserved sequence features in key bacteriochlorophyll biosynthesis proteins. Mol. Biol. Evol. *29*, 3397-3412.

Gupta, R.S. (2014). Identification of Conserved Indels that are Useful for Classification and Evolutionary Studies. In Methods in Microbiology New Approaches to Prokaryotics Systematics, Goodfellow M, Sutcliffe IC, and Chun J, eds. (London: Elsevier), pp. 153-182.

Gupta, R.S. (2016a). Editorial: Applications of Genome Sequences for Discovering Characteristics that Are Unique to Different Groups of Organisms and Provide Insights into Evolutionary Relationships. Front Genet. *7*, 27.

Gupta, R.S. (2016b). Impact of genomics on the understanding of microbial evolution and classification: the importance of Darwin's views on classification. FEMS Microbiol. Rev. *40*, 520-553.

Gupta, R.S. (2016c). Molecular signatures that are distinctive characteristics of the vertebrates and chordates and supporting a grouping of vertebrates with the tunicates. Mol. Phylogenet. Evol *94*, 383-391.

Gupta, R.S. (2018). Impact of Genomics on Clarifying the Evolutionary Relationships amongst Mycobacteria: Identification of Molecular Signatures Specific for the Tuberculosis-Complex of Bacteria with Potential Applications for Novel Diagnostics and Therapeutics. High Throughput. *7*.

Gupta, R.S. and Bhandari, V. (2011). Phylogeny and molecular signatures for the phylum Thermotogae and its subgroups. Antonie Van Leeuwenhoek *100*, 1-34.

Gupta, R.S. and Epand, R.M. (2017). Phylogenetic analysis of the diacylglycerol kinase family of proteins and identification of multiple highly-specific conserved inserts and deletions within the catalytic domain that are distinctive characteristics of different classes of DGK homologs. PLoS. One. *12*, e0182758.

Gupta, R.S. and Griffiths, E. (2002). Critical issues in bacterial phylogeny. Theor. Popul. Biol. *61*, 423-434.

Gupta, R.S. and Johari, V. (1998). Signature sequences in diverse proteins provide evidence of a close evolutionary relationship between the *Deinococcus-thermus* group and cyanobacteria. J. Mol. Evol. *46*, 716-720.

Gupta, R.S. and Khadka, B. (2015). Evidence for the presence of key chlorophyll-biosynthesis-related proteins in the genus *Rubrobacter* (Phylum Actinobacteria) and its implications for the evolution and origin of photosynthesis. Photosynth. Res. *127*, 201-218.

Gupta, R.S. and Lali, R. (2013). Molecular signatures for the phylum Aquificae and its different clades: proposal for division of the phylum Aquificae into the emended order *Aquificales*, containing the families *Aquificaceae* and *Hydrogenothermaceae*, and a new order *Desulfurobacteriales* ord. nov., containing the family *Desulfurobacteriaceae*. Antonie Van Leeuwenhoek *104*, 349-368.

Gupta, R.S., Nanda, A., and Khadka, B. (2017). Novel molecular, structural and evolutionary characteristics of the phosphoketolases from bifidobacteria and *Coriobacteriales*. PLoS. One. *12*, e0172176.

Gusev, O., Suetsugu, Y., Cornette, R., Kawashima, T., Logacheva, M.D., Kondrashov, A.S., Penin, A.A., Hatanaka, R., Kikuta, S., Shimura, S., Kanamori, H., Katayose, Y., Matsumoto, T., Shagimardanova, E., Alexeev, D., Govorun, V., Wisecaver, J., Mikheyev, A., Koyanagi, R., Fujie, M., Nishiyama, T., Shigenobu, S., Shibata, T.F., Golygina, V., Hasebe, M., Okuda, T., Satoh, N., and Kikawada, T. (2014). Comparative genome sequencing reveals genomic signature of extreme desiccation tolerance in the anhydrobiotic midge. Nat. Commun. *5*, 4784.

Hagen, J.B. (2000). The origins of bioinformatics. Nat. Rev. Genet. *1*, 231-236.

Hall, B.G. (2005). Comparison of the accuracies of several phylogenetic methods using protein and DNA sequences. Mol. Biol. Evol *22*, 792-802.

Hall, B.G. (2006). Simple and accurate estimation of ancestral protein sequences. Proc Natl. Acad. Sci. U. S. A *103*, 5431-5436.

Hall, J.E., Freites, J.A., and Tobias, D.J. (2019). Experimental and Simulation Studies of Aquaporin 0 Water Permeability and Regulation. Chem Rev. *119*, 6015-6039.

Halperin, I., Ma, B., Wolfson, H., and Nussinov, R. (2002). Principles of docking: An overview of search algorithms and a guide to scoring functions. Proteins *47*, 409-443.

Harris, J.I., Naughton, Ma., and Sanger (1956). Species differences in insulin. Arch Biochem Biophys. *65*, -427.

Hartley, B.S. (1970). Homologies in serine proteinases. Philos. Trans. R Soc Lond B Biol Sci. *257*, 77-87.

Hashimoto, K. and Panchenko, A.R. (2010). Mechanisms of protein oligomerization, the critical role of insertions and deletions in maintaining different oligomeric states. Proc. Natl. Acad. Sci. U. S. A *107*, 20352-20357.

Hassan, F.M.N. and Gupta, R.S. (2018). Novel Sequence Features of DNA Repair Genes/Proteins from *Deinococcus* Species Implicated in Protection from Oxidatively Generated Damage. Genes *9*, 149.

Hayakawa, N., Noguchi, M., Takeshita, S., Eviryanti, A., Seki, Y., Nishio, H., Yokoyama, R., Noguchi, M., Shuto, M., Shima, Y., Kuribayashi, K., Kageyama, S., Eda, H., Suzuki, M., Hatta, T., Iemura, S., Natsume, T., Tanabe, I., Nakagawa, R., Shiozaki, M., Sakurai, K., Shoji, M., Andou, A., and Yamamoto, T. (2014). Structure-activity relationship study, target identification, and pharmacological characterization of a small molecular IL-12/23 inhibitor, APY0201. Bioorg. Med. Chem *22*, 3021-3029.

Heather, J.M. and Chain, B. (2016). The sequence of sequencers: The history of sequencing DNA. Genomics *107*, 1-8.

Heck, J.N., Mellman, D.L., Ling, K., Sun, Y., Wagoner, M.P., Schill, N.J., and Anderson, R.A. (2007). A conspicuous connection: structure defines function for the phosphatidylinositol-phosphate kinase family. Crit Rev Biochem Mol. Biol *42*, 15-39.

Heilmann, I. (2016). Phosphoinositide signaling in plant development. Development *143*, 2044-2055.

Henard, C.A., Freed, E.F., and Guarnieri, M.T. (2015). Phosphoketolase pathway engineering for carbon-efficient biocatalysis. Curr. Opin. Biotechnol. *36*, 183-188.

Henzler-Wildman, K. and Kern, D. (2007). Dynamic personalities of proteins. Nature *450*, 964-972.

Hillisch, A., Pineda, L.F., and Hilgenfeld, R. (2004). Utility of homology models in the drug discovery process. Drug Discov. Today *9*, 659-669.

Ho, J., Adeolu, M., Khadka, B., and Gupta, R.S. (2016). Identification of distinctive molecular traits that are characteristic of the phylum "Deinococcus-Thermus" and distinguish its main constituent groups. Syst. Appl. Microbiol. *39*, 453-463.

Hoffmann, D., Kramer, B., Washio, T., Steinmetzer, T., Rarey, M., and Lengauer, T. (1999). Two-stage method for protein-ligand docking. J Med. Chem *42*, 4422-4433.

Hollingsworth, S.A. and Dror, R.O. (2018). Molecular Dynamics Simulation for All. Neuron *99*, 1129-1143.

Holyoake, J. and Sansom, M.S. (2007). Conformational change in an MFS protein: MD simulations of LacY. Structure. *15*, 873-884.

Homma, K., Terui, S., Minemura, M., Qadota, H., Anraku, Y., Kanaho, Y., and Ohya, Y. (1998). Phosphatidylinositol-4-phosphate 5-kinase localized on the plasma membrane is essential for yeast cell morphogenesis. J Biol. Chem *273*, 15779-15786.

Hood, L. and Rowen, L. (2013). The Human Genome Project: big science transforms biology and medicine. Genome Med. *5*, 79.

Hormozdiari, F., Salari, R., Hsing, M., Schonhuth, A., Chan, S.K., Sahinalp, S.C., and Cherkasov, A. (2009). The effect of insertions and deletions on wirings in protein-protein interaction networks: a large-scale study. J Comput. Biol *16*, 159-167.

Hsing, M. and Cherkasov, A. (2008). Indel PDB: a database of structural insertions and deletions derived from sequence alignments of closely related proteins. BMC. Bioinformatics *9*, 293.

Huang, N. and Jacobson, M.P. (2007). Physics-based methods for studying protein-ligand interactions. Curr. Opin. Drug Discov. Devel. *10*, 325-331.

Huelsenbeck, J.P., Ronquist, F., Nielsen, R., and Bollback, J.P. (2001). Bayesian inference of phylogeny and its impact on evolutionary biology. Science *294*, 2310-2314.

Ilari, A. and Savino, C. (2008). Protein structure determination by x-ray crystallography. Methods Mol. Biol. *452*, 63-87.

Ishihara, H., Shibasaki, Y., Kizuki, N., Katagiri, H., Yazaki, Y., Asano, T., and Oka, Y. (1996). Cloning of cDNAs encoding two isoforms of 68-kDa type I phosphatidylinositol-4-phosphate 5-kinase. J. Biol. Chem. *271*, 23611-23614.

Ishihara, H., Shibasaki, Y., Kizuki, N., Wada, T., Yazaki, Y., Asano, T., and Oka, Y. (1998). Type I phosphatidylinositol-4-phosphate 5-kinases. Cloning of the third isoform and deletion/substitution analysis of members of this novel lipid kinase family. J. Biol. Chem. *273*, 8741-8748.

Jack, B.R., Meyer, A.G., Echave, J., and Wilke, C.O. (2016). Functional Sites Induce Long-Range Evolutionary Constraints in Enzymes. PLoS. Biol *14*, e1002452.

Jain, M., Koren, S., Miga, K.H., Quick, J., Rand, A.C., Sasani, T.A., Tyson, J.R., Beggs, A.D., Dilthey, A.T., Fiddes, I.T., Malla, S., Marriott, H., Nieto, T., O'Grady, J., Olsen, H.E., Pedersen, B.S., Rhie, A., Richardson, H., Quinlan, A.R., Snutch, T.P., Tee, L., Paten, B., Phillippy, A.M., Simpson, J.T., Loman, N.J., and Loose, M. (2018). Nanopore sequencing and assembly of a human genome with ultra-long reads. Nat. Biotechnol. *36*, 338-345.

Janecka, J.E., Miller, W., Pringle, T.H., Wiens, F., Zitzmann, A., Helgen, K.M., Springer, M.S., and Murphy, W.J. (2007). Molecular and genomic data identify the closest living relative of primates. Science *318*, 792-794.

Janin, J. (2010). Protein-protein docking tested in blind predictions: the CAPRI experiment. Mol. Biosyst. *6*, 2351-2362.

Jaroszewski, L. (2009). Protein structure prediction based on sequence similarity. Methods Mol. Biol. *569*, 129-156.

Kalli, A.C., Devaney, I., and Sansom, M.S. (2014). Interactions of phosphatase and tensin homologue (PTEN) proteins with phosphatidylinositol phosphates: insights from molecular dynamics simulations of PTEN and voltage sensitive phosphatase. Biochemistry *53*, 1724-1732.

Kalli, A.C. and Sansom, M.S. (2014). Interactions of peripheral proteins with model membranes as viewed by molecular dynamics simulations. Biochem. Soc Trans. *42*, 1418-1424.

Karlin, S., Weinstock, G.M., and Brendel, V. (1995). Bacterial classifications derived from recA protein sequence comparisons. J Bacteriol. *177*, 6881-6893.

Karplus, M. (2002). Molecular dynamics simulations of biomolecules. Acc. Chem. Res. *35*, 321-323.

Karplus, M. and Kuriyan, J. (2005). Molecular dynamics and protein function. Proc. Natl. Acad. Sci. U. S. A *102*, 6679-6685.

Kartha, G., Bello, J., and Harker, D. (1967). Tertiary structure of ribonuclease. Nature *213*, 862-865.

Katoh, K., Kuma, K., Miyata, T., and Toh, H. (2005). Improvement in the accuracy of multiple sequence alignment program MAFFT. Genome Inform. *16*, 22-33.

Ke, R., Mignardi, M., Hauling, T., and Nilsson, M. (2016). Fourth Generation of Next-Generation Sequencing Technologies: Promise and Consequences. Hum. Mutat. *37*, 1363-1367.

Kelley, L.A., Mezulis, S., Yates, C.M., Wass, M.N., and Sternberg, M.J. (2015). The Phyre2 web portal for protein modeling, prediction and analysis. Nat. Protoc. *10*, 845-858.

Kendrew, J.C., Bodo, G., Dintizis, H.M., Parrish, R.G., Wyckoff, H., and Phillips, D.C. (1958). A three-dimensional model of the myoglobin molecule obtained by x-ray analysis. Nature *181*, 662-666.

Khadka, B., Adeolu, M., Blankenship, R.E., and Gupta, R.S. (2017). Novel insights into the origin and diversification of photosynthesis based on analyses of conserved indels in the core reaction center proteins. Photosynth. Res. *131*, 159-171.

Khadka, B., Chatterjee, T., Gupta, B.P., and Gupta, R.S. (2019). Genomic Analyses Identify Novel Molecular Signatures Specific for the *Caenorhabditis* and other Nematode Taxa Providing Novel Means for Genetic and Biochemical Studies. Genes. *10*, 739.

Khadka, B. and Gupta, R.S. (2017). Identification of a conserved 8 aa insert in the PIP5K protein in the *Saccharomycetaceae* family of fungi and the molecular dynamics simulations and structural analysis to investigate its potential functional role. Proteins *85*, 1454-1467.

Khadka, B. and Gupta, R.S. (2019). Novel Molecular Signatures in the PIP4K/PIP5K Family of Proteins Specific for Different Isozymes and Subfamilies Provide Important Insights into the Evolutionary Divergence of this Protein Family. Genes *10*, 312.

Khafizov, K., Madrid-Aliste, C., Almo, S.C., and Fiser, A. (2014). Trends in structural coverage of the protein universe and the impact of the Protein Structure Initiative. Proc. Natl. Acad. Sci. U. S. A *111*, 3733-3738.

Kim, R. and Guo, J.T. (2010). Systematic analysis of short internal indels and their impact on protein folding. BMC. Struct. Biol *10*, 24.

Kinoshita, K. and Nakamura, H. (2003). Protein informatics towards function identification. Curr. Opin. Struct. Biol. *13*, 396-400.

Kleckner, I.R. and Foster, M.P. (2011). An introduction to NMR-based approaches for measuring protein dynamics. Biochim. Biophys. Acta *1814*, 942-968.

Klenk, H.P. and Goker, M. (2010). En route to a genome-based classification of Archaea and Bacteria? Syst. Appl. Microbiol. *33*, 175-182.

Klepeis, J.L., Lindorff-Larsen, K., Dror, R.O., and Shaw, D.E. (2009). Long-timescale molecular dynamics simulations of protein structure and function. Curr. Opin. Struct. Biol. *19*, 120-127.

Kleywegt, G.J. and Jones, T.A. (1996). Phi/psi-chology: Ramachandran revisited. Structure. *4*, 1395-1400.

Kneale, G.G. and Bishop, M.J. (1985). Nucleic acid and protein sequence databases. Comput. Appl. Biosci. *1*, 11-17.

Kneale, G.G. and Kennard, O. (1984). The EMBL nucleotide sequence data library. Biochem. Soc Trans. *12*, 1011-1014.

Koonin, E.V., Aravind, L., and Kondrashov, A.S. (2000). The impact of comparative genomics on our understanding of evolution. Cell *101*, 573-576.

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

Korzheva, N., Mustaev, A., Kozlov, M., Malhotra, A., Nikiforov, V., Goldfarb, A., and Darst, S.A. (2000). A structural model of transcription elongation. Science *289*, 619-625.

Kozakov, D., Brenke, R., Comeau, S.R., and Vajda, S. (2006). PIPER: an FFT-based protein docking program with pairwise potentials. Proteins *65*, 392-406.

Kristensen, D.M., Ward, R.M., Lisewski, A.M., Erdin, S., Chen, B.Y., Fofanov, V.Y., Kimmel, M., Kavraki, L.E., and Lichtarge, O. (2008). Prediction of enzyme function based on 3D templates of evolutionarily important amino acids. BMC. Bioinformatics *9*, 17.

Kryshtafovych, A., Fidelis, K., and Moult, J. (2014). CASP10 results compared to those of previous CASP experiments. Proteins *82 Suppl 2*, 164-174.

Kullmann, W. (1991). Aristotle as a Natural Scientist. Acta Classica ***34***, 150.

Kundrotas, P.J., Zhu, Z., Janin, J., and Vakser, I.A. (2012). Templates are available to model nearly all complexes of structurally characterized proteins. Proc. Natl. Acad. Sci. U. S. A *109*, 9438-9441.

Kuntz, I.D., Blaney, J.M., Oatley, S.J., Langridge, R., and Ferrin, T.E. (1982). A geometric approach to macromolecule-ligand interactions. J Mol. Biol. *161*, 269-288.

Kutateladze, T.G. (2010). Translation of the phosphoinositide code by PI effectors. Nat. Chem. Biol. *6*, 507-513.

Kutzner, C., Pall, S., Fechner, M., Esztermann, A., de Groot, B.L., and Grubmuller, H. (2015). Best bang for your buck: GPU nodes for GROMACS biomolecular simulations. J Comput. Chem *36*, 1990-2008.

Kuznedelov, K., Minakhin, L., Niedziela-Majka, A., Dove, S.L., Rogulja, D., Nickels, B.E., Hochschild, A., Heyduk, T., and Severinov, K. (2002). A role for interaction of the RNA polymerase flap domain with the sigma subunit in promoter recognition. Science *295*, 855-857.

Kyrpides, N.C. (2009). Fifteen years of microbial genomics: meeting the challenges and fulfilling the dream. Nat. Biotechnol. *27*, 627-632.

Landick, R., Stewart, J., and Lee, D.N. (1990). Amino acid changes in conserved regions of the beta-subunit of *Escherichia coli* RNA polymerase alter transcription pausing and termination. Genes Dev. *4*, 1623-1636.

Lane, W.J. and Darst, S.A. (2010). Molecular evolution of multisubunit RNA polymerases: sequence analysis. J. Mol. Biol. *395*, 671-685.

Larsson, P., Hess, B., and Lindahl, E. (2019). Algorithm improvements for molecular dynamics simulations. Wiley Interdisciplinary Reviews: Computational Molecular Science *1*, 93-108.

Launay, G. and Simonson, T. (2008). Homology modelling of protein-protein complexes: a simple method and its possibilities and limitations. BMC. Bioinformatics. *9*, 427.

Lee, D., Redfern, O., and Orengo, C. (2007). Predicting protein function from sequence and structure. Nat. Rev. Mol. Cell Biol. *8*, 995-1005.

Lee, E.H., Hsin, J., Sotomayor, M., Comellas, G., and Schulten, K. (2009). Discovery through the computational microscope. Structure. *17*, 1295-1306.

Lensink, M.F., Velankar, S., and Wodak, S.J. (2017). Modeling protein-protein and protein-peptide complexes: CAPRI 6th edition. Proteins *85*, 359-377.

Lesk, A.M. and Chothia, C. (1980). How different amino acid sequences determine similar protein structures: the structure and evolutionary dynamics of the globins. J Mol. Biol *136*, 225-270.

Levinthal, C. (1966). Molecular model-building by computer. Sci. Am. *214*, 42-52.

Levinthal, C., Wodak, S.J., Kahn, P., and Dadivanian, A.K. (1975). Hemoglobin interaction in sickle cell fibers. I: Theoretical approaches to the molecular contacts. Proc. Natl. Acad. Sci. U. S. A *72*, 1330-1334.

Levit, A., Barak, D., Behrens, M., Meyerhof, W., and Niv, M.Y. (2012). Homology model-assisted elucidation of binding sites in GPCRs. Methods Mol. Biol. *914*, 179-205.

Levitt, M. and Warshel, A. (1975). Computer simulation of protein folding. Nature *253*, 694-698.

Lewis, D.F. (1999). Homology modelling of human cytochromes P450 involved in xenobiotic metabolism and rationalization of substrate selectivity. Exp. Toxicol. Pathol. *51*, 369-374.

Lewis, D.F., Dickins, M., Lake, B.G., Eddershaw, P.J., Tarbit, M.H., and Goldfarb, P.S. (1999). Molecular modelling of the human cytochrome P450 isoform CYP2A6 and investigations of CYP2A substrate selectivity. Toxicology *133*, 1-33.

Li, L., Liu, Z., Meng, D., Liu, X., Li, X., Zhang, M., Tao, J., Gu, Y., Zhong, S., and Yin, H. (2019). Comparative Genomic Analysis Reveals the Distribution, Organization, and Evolution of Metal Resistance Genes in the Genus *Acidithiobacillus*. Appl. Environ. Microbiol. *85*.

Lindorff-Larsen, K., Maragakis, P., Piana, S., Eastwood, M.P., Dror, R.O., and Shaw, D.E. (2012). Systematic validation of protein force fields against experimental data. PLoS. One. *7*, e32131.

Linneaus, C. (1758). Systema Naturae, edition X, Vol. 1 (Systema naturae per regna tria naturae, secundum classes, ordines, genera, species, cum characteribus, differentiis, synonymis, locis.) Editio Decima, Reformata. Tomus I. Laurentii Salvii, Stockholm. 824 pp.

Lipscomb, W.N., Hartsuck, J.A., Quiocho, F.A., and Reeke, G.N., Jr. (1969). The structure of carboxypeptidase A. IX. The x-ray diffraction results in the light of the chemical sequence. Proc Natl. Acad. Sci. U. S. A *64*, 28-35.

Liu, J. and Wang, R. (2015). Classification of current scoring functions. J Chem Inf. Model. *55*, 475-482.

Loijens, J.C., Boronenkov, I.V., Parker, G.J., and Anderson, R.A. (1996). The phosphatidylinositol 4-phosphate 5-kinase family. Adv. Enzyme Regul. *36*, 115-140.

Lopez, M., Cherkasov, A., and Nandan, D. (2007). Molecular architecture of leishmania EF-1alpha reveals a novel site that may modulate protein translation: a possible target for drug development. Biochem. Biophys Res. Commun. *356*, 886-892.

Lovell, S.C., Davis, I.W., Arendall, W.B., III, de Bakker, P.I., Word, J.M., Prisant, M.G., Richardson, J.S., and Richardson, D.C. (2003). Structure validation by Calpha geometry: phi,psi and Cbeta deviation. Proteins *50*, 437-450.

Lukk, T., Sakai, A., Kalyanaraman, C., Brown, S.D., Imker, H.J., Song, L., Fedorov, A.A., Fedorov, E.V., Toro, R., Hillerich, B., Seidel, R., Patskovsky, Y., Vetting, M.W., Nair, S.K., Babbitt, P.C., Almo, S.C., Gerlt, J.A., and Jacobson, M.P. (2012). Homology models guide discovery of diverse enzyme specificities among dipeptide epimerases in the enolase superfamily. Proc. Natl. Acad. Sci. U. S. A *109*, 4122-4127.

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

Luttmann, E., Ensign, D.L., Vaidyanathan, V., Houston, M., Rimon, N., Oland, J., Jayachandran, G., Friedrichs, M., and Pande, V.S. (2009). Accelerating molecular dynamic simulation on the cell processor and Playstation 3. J Comput. Chem *30*, 268-274.

Lyskov, S., Chou, F.C., Conchuir, S.O., Der, B.S., Drew, K., Kuroda, D., Xu, J., Weitzner, B.D., Renfrew, P.D., Sripakdeevong, P., Borgo, B., Havranek, J.J., Kuhlman, B., Kortemme, T., Bonneau, R., Gray, J.J., and Das, R. (2013). Serverification of molecular modeling applications: the Rosetta Online Server that Includes Everyone (ROSIE). PLoS. One. *8*, e63906.

Lyskov, S. and Gray, J.J. (2008). The RosettaDock server for local protein-protein docking. Nucleic Acids Res. *36*, W233-W238.

Ma, J., Sigler, P.B., Xu, Z., and Karplus, M. (2000). A dynamic model for the allosteric mechanism of GroEL. J Mol. Biol. *302*, 303-313.

Majerus, P.W. (1992). Inositol phosphate biochemistry. Annu. Rev. Biochem. *61*, 225-250.

Margoliash, E., FROHWIRT, N., and WIENER, E. (1959). A study of the cytochrome c haemochromogen. Biochem J *71*, 559-570.

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

Martin, T.F. (1998). Phosphoinositide lipids as signaling molecules: common themes for signal transduction, cytoskeletal regulation, and membrane trafficking. Annu Rev Cell Dev Biol. *14*, 231-264.

Mashimo, T., Fukunishi, Y., Kamiya, N., Takano, Y., Fukuda, I., and Nakamura, H. (2013). Molecular Dynamics Simulations Accelerated by GPU for Biological Macromolecules with a Non-Ewald Scheme for Electrostatic Interactions. J Chem Theory. Comput. *9*, 5599-5609.

Matsuura, T., Miyai, K., Trakulnaleamsai, S., Yomo, T., Shima, Y., Miki, S., Yamamoto, K., and Urabe, I. (1999). Evolutionary molecular engineering by random elongation mutagenesis. Nat. Biotechnol. *17*, 58-61.

Maxam, A.M. and Gilbert, W. (1977). A new method for sequencing DNA. Proc. Natl. Acad. Sci. U. S. A *74*, 560-564.

Mayr, E. (1982). The Growth of Biological Thought. Diversity, Evolution, and Inheritance. Harvard University Press.

McCammon, J.A., Gelin, B.R., and Karplus, M. (1977). Dynamics of folded proteins. Nature *267*, 585-590.

McInnes, C. (2007). Virtual screening strategies in drug discovery. Curr. Opin. Chem Biol. *11*, 494-502.

McPherson, J.D. (2014). A defining decade in DNA sequencing. Nat. Methods *11*, 1003-1005.

Medini, D., Donati, C., Tettelin, H., Masignani, V., and Rappuoli, R. (2005). The microbial pan-genome. Curr. Opin. Genet. Dev. *15*, 589-594.

Meile, L., Rohr, L.M., Geissmann, T.A., Herensperger, M., and Teuber, M. (2001). Characterization of the D-xylulose 5-phosphate/D-fructose 6-phosphate phosphoketolase gene (xfp) from *Bifidobacterium lactis*. J Bacteriol. *183*, 2929-2936.

Meng, X.Y., Zhang, H.X., Mezei, M., and Cui, M. (2011). Molecular docking: a powerful approach for structure-based drug discovery. Curr. Comput. Aided Drug Des *7*, 146-157.

Messing, J., Crea, R., and Seeburg, P.H. (1981). A system for shotgun DNA sequencing. Nucleic Acids Res. *9*, 309-321.

Milenkovic, S. and Bondar, A.N. (2016). Mechanism of conformational coupling in SecA: Key role of hydrogen-bonding networks and water interactions. Biochim Biophys Acta *1858*, 374-385.

Minakhin, L., Bhagat, S., Brunning, A., Campbell, E.A., Darst, S.A., Ebright, R.H., and Severinov, K. (2001). Bacterial RNA polymerase subunit omega and eukaryotic RNA polymerase subunit RPB6 are sequence, structural, and functional homologs and promote RNA polymerase assembly. Proc. Natl. Acad. Sci. U. S. A *98*, 892-897.

Mistry, J., Kloppmann, E., Rost, B., and Punta, M. (2013). An estimated 5% of new protein structures solved today represent a new Pfam family. Acta Crystallogr. D. Biol. Crystallogr. *69*, 2186-2193.

Mittermaier, A. and Kay, L.E. (2006). New tools provide new insights in NMR studies of protein dynamics. Science *312*, 224-228.

Moreira, I.S., Fernandes, P.A., and Ramos, M.J. (2010). Protein-protein docking dealing with the unknown. J Comput. Chem *31*, 317-342.

Mori, H. and Ito, K. (2001). The Sec protein-translocation pathway. Trends Microbiol. *9*, 494-500.

Morris, G.M., Huey, R., Lindstrom, W., Sanner, M.F., Belew, R.K., Goodsell, D.S., and Olson, A.J. (2009). AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility. J. Comput. Chem. *30*, 2785-2791.

Muegge, I. (2002). A knowledge-based scoring function for protein-ligand interactions: Probing the reference state. In Virtual Screening: An Alternative or Complement to High Throughput Screening?, G.Klebe, ed. (Dordrecht: Springer Netherlands), pp. 99-114.

Mustaev, A., Kashlev, M., Zaychikov, E., Grachev, M., and Goldfarb, A. (1993). Active center rearrangement in RNA polymerase initiation complex. J. Biol. Chem. *268*, 19185-19187.

Nandan, D., Lopez, M., Ban, F., Huang, M., Li, Y., Reiner, N.E., and Cherkasov, A. (2007). Indel-based targeting of essential proteins in human pathogens that have close host orthologue(s): discovery of selective inhibitors for *Leishmania donovani* elongation factor-1alpha. Proteins *67*, 53-64.

Naushad, S., Adeolu, M., Goel, N., Khadka, B., Al Dahwi, A., and Gupta, R.S. (2015). Phylogenomic and molecular demarcation of the core members of the polyphyletic *pasteurellaceae* genera *actinobacillus*, *haemophilus*, and *pasteurella*. Int. J Genomics *2015*, 198560.

Needleman, S.B. and Blair, T.T. (1969). Homology of *Pseudomonas* cytochrome c-551 with eukaryotic c-cytochromes. Proc. Natl. Acad. Sci. U. S. A *63*, 1227-1233.

Nerenberg, P.S. and Head-Gordon, T. (2018). New developments in force fields for biomolecular simulations. Curr. Opin. Struct. Biol. *49*, 129-138.

Ngo, J.C., Huang, M., Roth, D.A., Furie, B.C., and Furie, B. (2008). Crystal structure of human factor VIII: implications for the formation of the factor IXa-factor VIIIa complex. Structure. *16*, 597-606.

Nguyen, C.T., Tanaka, K., Cao, Y., Cho, S.H., Xu, D., and Stacey, G. (2016). Computational Analysis of the Ligand Binding Site of the Extracellular ATP Receptor, DORN1. PLoS. One. *11*, e0161894.

Notredame, C., Higgins, D.G., and Heringa, J. (2000). T-Coffee: A novel method for fast and accurate multiple sequence alignment. J Mol. Biol. *302*, 205-217.

Oda, T., Lim, K., and Tomii, K. (2017). Simple adjustment of the sequence weight algorithm remarkably enhances PSI-BLAST performance. BMC. Bioinformatics *18*, 288.

Ogden, T.H. and Rosenberg, M.S. (2006). Multiple sequence alignment accuracy and phylogenetic inference. Syst. Biol. *55*, 314-328.

Okazaki, K., Miyagishima, S.Y., and Wada, H. (2015). Phosphatidylinositol 4-phosphate negatively regulates chloroplast division in Arabidopsis. Plant Cell *27*, 663-674.

Oren, A. (2010). Concepts about Phylogeny of Microorganisms-an Historical Overview. In: Oren A, Papke RT (eds). *Molecular Phylogeny of Microorganisms*. (Norfolk: Caister Academic Press), pp. 1-21.

Oren, A. and Garrity, G.M. (2014). Then and now: a systematic review of the systematics of prokaryotes in the last 80 years. Antonie Van Leeuwenhoek *106*, 43-56.

Ostermeier, C. and Michel, H. (1997). Crystallization of membrane proteins. Curr. Opin. Struct. Biol. *7*, 697-701.

Oude Weernink, P.A., Schulte, P., Guo, Y., Wetzel, J., Amano, M., Kaibuchi, K., Haverland, S., Voss, M., Schmidt, M., Mayr, G.W., and Jakobs, K.H. (2000). Stimulation of phosphatidylinositol-4-phosphate 5-kinase by Rho-kinase. J Biol. Chem *275*, 10168-10174.

Overington, J., Donnelly, D., Johnson, M.S., Sali, A., and Blundell, T.L. (1992). Environment-specific amino acid substitution tables: tertiary templates and prediction of protein folds. Protein Sci. *1*, 216-226.

Park, H., Hwang, K.Y., Oh, K.H., Kim, Y.H., Lee, J.Y., and Kim, K. (2008). Discovery of novel alpha-glucosidase inhibitors based on the virtual screening with the homology-modeled protein structure. Bioorg. Med. Chem. *16*, 284-292.

Parkhill, J. and Wren, B.W. (2011). Bacterial epidemiology and biology--lessons from genome sequencing. Genome Biol. *12*, 230.

Pascarella, S. and Argos, P. (1992). Analysis of insertions/deletions in protein structures. J. Mol. Biol. *224*, 461-471.

Pauling, L., COREY, R.B., and BRANSON, H.R. (1951). The structure of proteins; two hydrogen-bonded helical configurations of the polypeptide chain. Proc Natl. Acad. Sci. U. S. A *37*, 205-211.

Payne, D.J., Gwynn, M.N., Holmes, D.J., and Pompliano, D.L. (2007). Drugs for bad bugs: confronting the challenges of antibacterial discovery. Nat. Rev. Drug Discov. *6*, 29-40.

Pellegrini, M., Marcotte, E.M., Thompson, M.J., Eisenberg, D., and Yeates, T.O. (1999). Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. Proc. Natl. Acad. Sci. U. S. A *96*, 4285-4288.

Perutz, M. (1985). Early days of protein crystallography. Methods Enzymol. *114*, 3-18.

Perutz, M.F. (1983). Species adaptation in a protein molecule. Mol. Biol Evol *1*, 1-28.

Perutz, M.F., Rossmann, M.G., Cullis, A.F., Muirhead, H., Will, G., and North, A.C. (1960). Structure of haemoglobin: a three-dimensional Fourier synthesis at 5.5-A. resolution, obtained by X-ray analysis. Nature *185*, 416-422.

Pettersen, E.F., Goddard, T.D., Huang, C.C., Couch, G.S., Greenblatt, D.M., Meng, E.C., and Ferrin, T.E. (2004). UCSF Chimera--a visualization system for exploratory research and analysis. J Comput. Chem *25*, 1605-1612.

Philippon, H., Brochier-Armanet, C., and Perriere, G. (2015). Evolutionary history of phosphatidylinositol- 3-kinases: ancestral origin in eukaryotes and complex duplication patterns. BMC. Evol Biol. *15*, 226.

Phillips, J.C., Braun, R., Wang, W., Gumbart, J., Tajkhorshid, E., Villa, E., Chipot, C., Skeel, R.D., Kale, L., and Schulten, K. (2005). Scalable molecular dynamics with NAMD. J Comput. Chem *26*, 1781-1802.

Pierce, B.G., Hourai, Y., and Weng, Z. (2011). Accelerating protein docking in ZDOCK using an advanced 3D convolution library. PLoS. One. *6*, e24657.

Plimpton, S. (1995). Fast Parallel Algorithms for Short-Range Molecular Dynamics. J Comp Phys *117*, 1-19.

Poger, D., Caron, B., and Mark, A.E. (2016). Validating lipid force fields against experimental data: Progress, challenges and perspectives. Biochim Biophys Acta *1858*, 1556-1565.

Pokusaeva, K., Fitzgerald, G.F., and van Sinderen, D. (2011). Carbohydrate metabolism in Bifidobacteria. Genes Nutr. *6*, 285-306.

Protein Data Bank (1971). Protein Data Bank. Nature New Biol *233*, 233.

Ragan, M.A. (2009). Trees and networks before and after Darwin. Biol Direct. *4*, 43.

Ramasamy, D., Mishra, A.K., Lagier, J.C., Padhmanabhan, R., Rossi, M., Sentausa, E., Raoult, D., and Fournier, P.E. (2014). A polyphasic strategy incorporating genomic data for the taxonomic description of novel bacterial species. Int. J Syst. Evol Microbiol. *64*, 384-391.

Rao, N.A., Talwar, R., and Savithri, H.S. (2000). Molecular organization, catalytic mechanism and function of serine hydroxymethyltransferase--a potential target for cancer chemotherapy. Int. J Biochem. Cell Biol. *32*, 405-416.

Ravenhall, M., Skunca, N., Lassalle, F., and Dessimoz, C. (2015). Inferring horizontal gene transfer. PLoS. Comput. Biol. *11*, e1004095.

Reeves, G.A., Dallman, T.J., Redfern, O.C., Akpor, A., and Orengo, C.A. (2006). Structural diversity of domain superfamilies in the CATH database. J Mol. Biol. *360*, 725-741.

Ritchie, D.W. (2008). Recent progress and future directions in protein-protein docking. Curr. Protein Pept. Sci. *9*, 1-15.

Rivera, M.C. and Lake, J.A. (1992). Evidence that eukaryotes and eocyte prokaryotes are immediate relatives. Science *257*, 74-76.

Rokas, A. and Holland, P.W. (2000). Rare genomic changes as a tool for phylogenetics. Trends Ecol. Evol *15*, 454-459.

Rose, P.W., Bi, C., Bluhm, W.F., Christie, C.H., Dimitropoulos, D., Dutta, S., Green, R.K., Goodsell, D.S., Prlic, A., Quesada, M., Quinn, G.B., Ramos, A.G., Westbrook, J.D., Young, J., Zardecki, C., Berman, H.M., and Bourne, P.E. (2013). The RCSB Protein Data Bank: new resources for research and education. Nucleic Acids Res. *41*, D475-D482.

Rose, P.W., Prlic, A., Bi, C., Bluhm, W.F., Christie, C.H., Dutta, S., Green, R.K., Goodsell, D.S., Westbrook, J.D., Woo, J., Young, J., Zardecki, C., Berman, H.M., Bourne, P.E., and Burley, S.K. (2015). The RCSB Protein Data Bank: views of

structural biology for basic and applied research and education. Nucleic Acids Res. *43*, D345-D356.

Rudling, A., Orro, A., and Carlsson, J. (2018). Prediction of Ordered Water Molecules in Protein Binding Sites from Molecular Dynamics Simulations: The Impact of Ligand Binding on Hydration Networks. J Chem Inf. Model. *58*, 350-361.

Russell, R.B., Saqi, M.A., Sayle, R.A., Bates, P.A., and Sternberg, M.J. (1997). Recognition of analogous and homologous protein folds: analysis of sequence and structure conservation. J Mol. Biol *269*, 423-439.

Sahoo, B.R., Maharana, J., Bhoi, G.K., Lenka, S.K., Patra, M.C., Dikhit, M.R., Dubey, P.K., Pradhan, S.K., and Behera, B.K. (2014). A conformational analysis of mouse Nalp3 domain structures by molecular dynamics simulations, and binding site analysis. Mol. Biosyst. *10*, 1104-1116.

Saibil, H.R. (2000). Macromolecular structure determination by cryo-electron microscopy. Acta Crystallogr. D. Biol. Crystallogr. *56*, 1215-1222.

Saitou, N. and Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol. Biol. Evol *4*, 406-425.

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

Sali, A., Overington, J.P., Johnson, M.S., and Blundell, T.L. (1990). From comparisons of protein sequences and structures to protein modelling and design. Trends Biochem Sci. *15*, 235-240.

Sanchez, B., Delgado, S., Blanco-Miguez, A., Lourenco, A., Gueimonde, M., and Margolles, A. (2017). Probiotics, gut microbiota, and their influence on host health and disease. Mol. Nutr. Food Res. *61*.

Sanchez, R. and Sali, A. (1997). Advances in comparative protein-structure modelling. Curr. Opin. Struct. Biol. *7*, 206-214.

Sanderson, M.J., Purvis, A., and Henze, C. (1998). Phylogenetic supertrees: Assembling the trees of life. Trends Ecol. Evol *13*, 105-109.

Sanger, F. (1949). Species differences in insulins. Nature *164*, 529.

Sanger, F. (1959). Chemistry of insulin; determination of the structure of insulin opens the way to greater understanding of life processes. Science *129*, 1340-1344.

Sanger, F., Coulson, A.R., Hong, G.F., Hill, D.F., and Petersen, G.B. (1982). Nucleotide sequence of bacteriophage lambda DNA. J Mol. Biol. *162*, 729-773.

Sanger, F., Nicklen, S., and Coulson, A.R. (1977). DNA sequencing with chain-terminating inhibitors. Proc. Natl. Acad. Sci. U. S. A *74*, 5463-5467.

Sanger, F. and Tuppy, H. (1951). The amino-acid sequence in the phenylalanyl chain of insulin. I. The identification of lower peptides from partial hydrolysates. Biochem J *49*, 463-481.

Sapp, J. (2009). The New Foundations of Evolution: On the Tree of Life. (London: Oxford University Press).

Schadt, E.E., Turner, S., and Kasarskis, A. (2010). A window into third-generation sequencing. Hum. Mol. Genet. *19*, R227-R240.

Schaffer, A.A., Aravind, L., Madden, T.L., Shavirin, S., Spouge, J.L., Wolf, Y.I., Koonin, E.V., and Altschul, S.F. (2001). Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements. Nucleic Acids Res. *29*, 2994-3005.

Schloss, J.A. (2008). How to get genomes at one ten-thousandth the cost. Nat. Biotechnol. *26*, 1113-1115.

Schmidt, M.G. and Kiser, K.B. (1999). SecA: the ubiquitous component of preprotein translocase in prokaryotes. Microbes. Infect. *1*, 993-1004.

Schmidt, T., Bergner, A., and Schwede, T. (2014). Modelling three-dimensional protein structures for applications in drug design. Drug Discov. Today *19*, 890-897.

Schneidman-Duhovny, D., Inbar, Y., Nussinov, R., and Wolfson, H.J. (2005). PatchDock and SymmDock: servers for rigid and symmetric docking. Nucleic Acids Res. *33*, W363-W367.

Schoeffler, A.J., May, A.P., and Berger, J.M. (2010). A domain insertion in *Escherichia coli* GyrB adopts a novel fold that plays a critical role in gyrase function. Nucleic Acids Res. *38*, 7830-7844.

Schwede, T. (2013). Protein modeling: what happened to the "protein structure gap"? Structure. *21*, 1531-1540.

Schwede, T., Kopp, J., Guex, N., and Peitsch, M.C. (2003). SWISS-MODEL: An automated protein homology-modeling server. Nucleic Acids Res. *31*, 3381-3385.

Semenas, J., Hedblom, A., Miftakhova, R.R., Sarwar, M., Larsson, R., Shcherbina, L., Johansson, M.E., Harkonen, P., Sterner, O., and Persson, J.L. (2014). The role of PI3K/AKT-related PIP5K1alpha and the discovery of its selective inhibitor for treatment of advanced prostate cancer. Proc. Natl. Acad. Sci. U. S. A *111*, E3689-E3698.

Severinov, K., Mustaev, A., Kukarin, A., Muzzin, O., Bass, I., Darst, S.A., and Goldfarb, A. (1996). Structural modules of the large subunits of RNA polymerase. Introducing archaebacterial and chloroplast split sites in the beta and beta' subunits of *Escherichia coli* RNA polymerase. J. Biol. Chem. *271*, 27969-27974.

Sharma, H., Cheng, X., and Buolamwini, J.K. (2012). Homology model-guided 3D-QSAR studies of HIV-1 integrase inhibitors. J. Chem. Inf. Model. *52*, 515-544.

Shendure, J., Balasubramanian, S., Church, G.M., Gilbert, W., Rogers, J., Schloss, J.A., and Waterston, R.H. (2017). DNA sequencing at 40: past, present and future. Nature *550*, 345-353.

Shendure, J., Findlay, G.M., and Snyder, M.W. (2019). Genomic Medicine-Progress, Pitfalls, and Promise. Cell *177*, 45-57.

Shendure, J. and Ji, H. (2008). Next-generation DNA sequencing. Nat. Biotechnol. *26*, 1135-1145.

Shivakumar, D., Williams, J., Wu, Y., Damm, W., Shelley, J., and Sherman, W. (2010). Prediction of Absolute Solvation Free Energies using Molecular Dynamics Free Energy Perturbation and the OPLS Force Field. J Chem Theory. Comput. *6*, 1509-1519.

Sievers, F., Wilm, A., Dineen, D., Gibson, T.J., Karplus, K., Li, W., Lopez, R., McWilliam, H., Remmert, M., Soding, J., Thompson, J.D., and Higgins, D.G. (2011). Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol. Syst. Biol. *7*, 539.

Singh, B. and Gupta, R.S. (2009). Conserved inserts in the Hsp60 (GroEL) and Hsp70 (DnaK) proteins are essential for cellular growth. Mol. Genet. Genomics *281*, 361-373.

Sippl, M.J. (1993). Recognition of errors in three-dimensional structures of proteins. Proteins *17*, 355-362.

Smith, L.M., Sanders, J.Z., Kaiser, R.J., Hughes, P., Dodd, C., Connell, C.R., Heiner, C., Kent, S.B., and Hood, L.E. (1986). Fluorescence detection in automated DNA sequence analysis. Nature *321*, 674-679.

Sneath, P. and Sokal, R. (1973). Numerical taxonomy: the principles and practice of numerical classification. (San Francisco: Freeman).

Snel, B., Bork, P., and Huynen, M.A. (1999). Genome phylogeny based on gene content. Nat. Genet. *21*, 108-110.

Sousa, S.F., Fernandes, P.A., and Ramos, M.J. (2006). Protein-ligand docking: current status and future challenges. Proteins *65*, 15-26.

Stackebrandt, E. and Schumann, P. (2006). Introduction to the taxonomy of actinobacteria. In The Prokaryotes. A Handbook on the Biology of Bacteria, Dworkin M, Falkow S, Rosenberg E, Schleifer KH, and Stackebrandt E, eds., pp. 297-321.

Staden, R. (1979). A strategy of DNA sequencing employing computer programs. Nucleic Acids Res. *6*, 2601-2610.

Stanier, R.Y., Doudoroff.M, and Adelberg, E.A. (1963). The Microbial World, 2nd edn. (Englewood Cliffs: Prentice-Hall).

Stanier, R.Y. and Van Niel, C.B. (1941). The Main Outlines of Bacterial Classification. J Bacteriol. *42*, 437-466.

Sutcliffe, I.C. (2015). Challenging the anthropocentric emphasis on phenotypic testing in prokaryotic species descriptions: rip it up and start again. Front Genet. *6*, 218.

Sutcliffe, M.J., Dobson, C.M., and Oswald, R.E. (1992). Solution structure of neuronal bungarotoxin determined by two-dimensional NMR spectroscopy: calculation of tertiary structure using systematic homologous model building, dynamical simulated annealing, and restrained molecular dynamics. Biochemistry *31*, 2962-2970.

Takahashi, K., Tagami, U., Shimba, N., Kashiwagi, T., Ishikawa, K., and Suzuki, E. (2010). Crystal structure of *Bifidobacterium longum* phosphoketolase; key enzyme for glucose metabolism in *Bifidobacterium*. FEBS Lett. *584*, 3855-3861.

Tang, C.M. and Moxon, E.R. (2001). The impact of microbial genomics on antimicrobial drug development. Annu. Rev. Genomics Hum. Genet. *2*, 259-269.

Tateno, Y., Takezaki, N., and Nei, M. (1994). Relative efficiencies of the maximum-likelihood, neighbor-joining, and maximum-parsimony methods when substitution rate varies with site. Mol. Biol. Evol *11*, 261-277.

Tatusov, R.L., Koonin, E.V., and Lipman, D.J. (1997). A genomic perspective on protein families. Science *278*, 631-637.

Terwilliger, T.C. (2011). The success of structural genomics. J. Struct. Funct. Genomics *12*, 43-44.

Tettelin, H., Masignani, V., Cieslewicz, M.J., Donati, C., Medini, D., Ward, N.L., Angiuoli, S.V., Crabtree, J., Jones, A.L., Durkin, A.S., Deboy, R.T., Davidsen, T.M., Mora, M., Scarselli, M., Ros, I., Peterson, J.D., Hauser, C.R., Sundaram,

J.P., Nelson, W.C., Madupu, R., Brinkac, L.M., Dodson, R.J., Rosovitz, M.J., Sullivan, S.A., Daugherty, S.C., Haft, D.H., Selengut, J., Gwinn, M.L., Zhou, L., Zafar, N., Khouri, H., Radune, D., Dimitrov, G., Watkins, K., O'Connor, K.J., Smith, S., Utterback, T.R., White, O., Rubens, C.E., Grandi, G., Madoff, L.C., Kasper, D.L., Telford, J.L., Wessels, M.R., Rappuoli, R., and Fraser, C.M. (2005). Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: implications for the microbial "pan-genome". Proc. Natl. Acad. Sci. U. S. A *102*, 13950-13955.

Tettelin, H., Riley, D., Cattuto, C., and Medini, D. (2008). Comparative genomics: the bacterial pan-genome. Curr. Opin. Microbiol. *11*, 472-477.

Thiel, K.A. (2004). Structure-aided drug design's next generation. Nat. Biotechnol. *22*, 513-519.

Trott, O. and Olson, A.J. (2010). AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. J. Comput. Chem. *31*, 455-461.

Tsilibaris, V., Maenhaut-Michel, G., and Van Melderen, L. (2006). Biological roles of the Lon ATP-dependent protease. Res. Microbiol. *157*, 701-713.

Turroni, F., van Sinderen, D., and Ventura, M. (2011). Genomics and ecological overview of the genus *Bifidobacterium*. Int. J Food Microbiol. *149*, 37-44.

Tyzack, J.D., Furnham, N., Sillitoe, I., Orengo, C.M., and Thornton, J.M. (2017). Understanding enzyme function evolution from a computational perspective. Curr. Opin. Struct. Biol. *47*, 131-139.

Ubarretxena-Belandia, I. and Stokes, D.L. (2010). Present and future of membrane protein structure determination by electron crystallography. Adv. Protein Chem. Struct. Biol. *81*, 33-60.

UniProt Consortium (2019). UniProt: a worldwide hub of protein knowledge. Nucleic Acids Res. *47*, D506-D515.

Vajda, S. and Camacho, C.J. (2004). Protein-protein docking: is the glass half-full or half-empty? Trends Biotechnol. *22*, 110-116.

Valas, R.E. and Bourne, P.E. (2009). Structural analysis of polarizing indels: an emerging consensus on the root of the tree of life. Biol. Direct. *4*, 30.

van den Bout, I. and Divecha, N. (2009). PIP5K-driven PtdIns(4,5)P2 synthesis: regulation and cellular functions. J. Cell Sci. *122*, 3837-3850.

Van Der, S.D., Lindahl, E., Hess, B., Groenhof, G., Mark, A.E., and Berendsen, H.J. (2005). GROMACS: fast, flexible, and free. J Comput. Chem *26*, 1701-1718.

van Dijk, E.L., Auger, H., Jaszczyszyn, Y., and Thermes, C. (2014). Ten years of next-generation sequencing technology. Trends Genet. *30*, 418-426.

van Dijk, E.L., Jaszczyszyn, Y., Naquin, D., and Thermes, C. (2018). The Third Revolution in Sequencing Technology. Trends Genet. *34*, 666-681.

Vattulainen, I. and Rog, T. (2011). Lipid simulations: a perspective on lipids in action. Cold Spring Harb. Perspect. Biol. *3*.

Vaughan, T.W. (1906). THE WORK OF HUGO DE VRIES AND ITS IMPORTANCE IN THE STUDY OF PROBLEMS OF EVOLUTION. Science *23*, 681-691.

Ventura, M., Turroni, F., Lugli, G.A., and van Sinderen, D. (2014). Bifidobacteria and humans: our special friends, from ecological to genomics perspectives. J Sci. Food Agric. *94*, 163-168.

Verma, M., Lal, D., Kaur, J., Saxena, A., Kaur, J., Anand, S., and Lal, R. (2013). Phylogenetic analyses of phylum Actinobacteria based on whole genome sequences. Res. Microbiol. *164*, 718-728.

Voss, M.D., Czechtizky, W., Li, Z., Rudolph, C., Petry, S., Brummerhop, H., Langer, T., Schiffer, A., and Schaefer, H.L. (2014). Discovery and pharmacological characterization of a novel small molecule inhibitor of phosphatidylinositol-5-phosphate 4-kinase, type II, beta. Biochem. Biophys Res. Commun. *449*, 327-331.

Vrontou, E. and Economou, A. (2004). Structure and function of SecA, the preprotein translocase nanomotor. Biochim Biophys Acta *1694*, 67-80.

Wacker, D., Wang, S., McCorvy, J.D., Betz, R.M., Venkatakrishnan, A.J., Levit, A., Lansu, K., Schools, Z.L., Che, T., Nichols, D.E., Shoichet, B.K., Dror, R.O., and Roth, B.L. (2017). Crystal Structure of an LSD-Bound Human Serotonin Receptor. Cell *168*, 377-389.

Walsh, C.T. and Wencewicz, T.A. (2014). Prospects for new antibiotics: a molecule-centered perspective. J. Antibiot. (Tokyo) *67*, 7-22.

Wang, Y., Geer, L.Y., Chappey, C., Kans, J.A., and Bryant, S.H. (2000). Cn3D: sequence and structure views for Entrez. Trends Biochem. Sci. *25*, 300-302.

Watson, J.D. and Crick, F.H. (1953). Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. Nature *171*, 737-738.

Watson, J.D., Laskowski, R.A., and Thornton, J.M. (2005). Predicting protein function from sequence and structural data. Curr. Opin. Struct. Biol. *15*, 275-284.

Wei, L., Liu, Y., Dubchak, I., Shon, J., and Park, J. (2002). Comparative genomics approaches to study organism similarities and differences. J Biomed. Inform. *35*, 142-150.

Whitman, W.B. (2015). Genome sequences as the type material for taxonomic descriptions of prokaryotes. Syst. Appl. Microbiol. *38*, 217-222.

Wiederstein, M. and Sippl, M.J. (2007). ProSA-web: interactive web service for the recognition of errors in three-dimensional structures of proteins. Nucleic Acids Res. *35*, W407-W410.

Wodak, S.J. and Janin, J. (1978). Computer analysis of protein-protein interaction. J Mol. Biol. *124*, 323-342.

Woese, C.R. (1987). Bacterial evolution. Microbiol. Rev. *51*, 221-271.

Woese, C.R. and Fox, G.E. (1977). Phylogenetic structure of the prokaryotic domain: the primary kingdoms. Proc Natl. Acad. Sci. U. S. A *74*, 5088-5090.

Woese, C.R., Kandler, O., and Wheelis, M.L. (1990). Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. Proc Natl. Acad. Sci. U. S. A *87*, 4576-4579.

Wong, S.Y., Paschos, A., Gupta, R.S., and Schellhorn, H.E. (2014). Insertion/Deletion-Based Approach for the Detection of *Escherichia coli* O157:H7 in Freshwater Environments. Environ. Sci. Technol. *48*, 11462-11470.

Wright, B.D., Simpson, C., Stashko, M., Kireev, D., Hull-Ryde, E.A., Zylka, M.J., and Janzen, W.P. (2015). Development of a High-Throughput Screening Assay to Identify Inhibitors of the Lipid Kinase PIP5K1C. J. Biomol. Screen. *20*, 655-662.

Wright, G. (2015). Antibiotics: An irresistible newcomer. Nature *517*, 442-444.

Xu, D. and Zhang, Y. (2011). Improving the physical realism and structural accuracy of protein models by a two-step atomic-level energy minimization. Biophys. J. *101*, 2525-2534.

Xu, D. and Zhang, Y. (2012). Ab initio protein structure assembly using continuous structure fragments and optimized knowledge-based force field. Proteins *80*, 1715-1735.

Xu, X., Song, H., Qi, J., Liu, Y., Wang, H., Su, C., Shi, Y., and Gao, G.F. (2016). Contribution of intertwined loop to membrane association revealed by Zika virus full-length NS1 structure. EMBO J *35*, 2170-2178.

Yamey, G. (2000). Scientists unveil first draft of human genome. BMJ *321*, 7.

Yang, J. and Zhang, Y. (2015). I-TASSER server: new development for protein structure and function predictions. Nucleic Acids Res. *43*, W174-W181.

Yin, X., Chambers, J.R., Barlow, K., Park, A.S., and Wheatcroft, R. (2005). The gene encoding xylulose-5-phosphate/fructose-6-phosphate phosphoketolase (*xfp*) is conserved among *Bifidobacterium* species within a more variable region of the genome and both are useful for strain identification. FEMS Microbiol. Lett. *246*, 251-257.

Zhang, G., Gao, B., Adeolu, M., Khadka, B., and Gupta, R.S. (2016a). Phylogenomic Analyses and Comparative Studies on Genomes of the *Bifidobacteriales*: Identification of Molecular Signatures Specific for the Order *Bifidobacteriales* and Its Different Subclades. Front Microbiol. *7*, 978.

Zhang, Q., Feng, T., Xu, L., Sun, H., Pan, P., Li, Y., Li, D., and Hou, T. (2016b). Recent Advances in Protein-Protein Docking. Curr. Drug Targets. *17*, 1586-1594.

Zhang, Z., Huang, J., Wang, Z., Wang, L., and Gao, P. (2011). Impact of indels on the flanking regions in structural domains. Mol. Biol. Evol *28*, 291-301.

Zhang, Z., Xing, C., Wang, L., Gong, B., and Liu, H. (2012). IndelFR: a database of indels in protein structures and their flanking regions. Nucleic Acids Res. *40*, D512-D518.

Zimmer, J. and Rapoport, T.A. (2009). Conformational flexibility and peptide interaction of the translocation ATPase SecA. J Mol. Biol. *394*, 606-612.

Zuckerkandl, E. and Pauling, L. (1965). Molecules as documents of evolutionary history. J Theor Biol. *8*, 357-366.