Exploring protein domain evolution by designing new TPR-like domains

der Fakultät für Biologie
der Eberhard Karls Universität Tübingen
zur Erlangung des Grades eines
Doktors der Naturwissenschaften

von

Manjunatha Karpenahalli Ranganathappa
aus Karpenahalli, Indien
vorgelegte

Dissertation

2006

Tag der mündlichen Prüfung: 15.12.2006

Dekan: Prof. Dr. Friedrich Schöffl

1. Berichterstatter: Prof. Dr. Alfred Nordheim

2. Berichterstatter: Prof. Dr. Andrei N. Lupas

Dedicated in the memory of my father, the late Sri Ranganathappa.

Acknowledgements

- I express my gratitude to Prof. Dr. Andrei Lupas, for sharing his vast knowledge and skills in various areas of science, and for his guidance throughout my PhD.
- I sincerely thank Dr. Jörg Martin whose expertise, understanding, and patience, added considerably to my graduate experience.
- I am very grateful to Prof. Dr. Alfred Nordheim for accepting me as a PhD student at the University of Tübingen.
- Prof. Dr. S. Krishna Swamy is the one professor/teacher who truly made a difference in my life. It was under his tutelage that I developed a focus and became interested in bioinformatics.
- A very special thanks to Dr. Ramadoss for helping me to strengthen my skills in biochemistry.
- My special thanks to Ms. Karin Lehmann for her assistance and support during my stay in Germany.
- I sincerely thank all my colleagues, the current and former members of the Protein Evolution Department for their co-operation and friendship during my stay in this department.
- I would also like to thank all my friends for exchange of knowledge and skills which helped enrich my PhD experience.
- I feel a deep sense of gratitude for my family members, who formed part of my vision and taught me the good things that really matter in life.

Abstract

Proteins are the most abundant and diverse class of biomolecules that mediate the vast majority of biochemical processes. The functional units within a protein are the ”domains” which fold autonomously from the rest of the linear amino acid sequence in the protein. Novelty in protein function often arises as a result of gain, loss or re-shuffling of existing domains. Thus, protein domains can arguably be seen as stable units of evolution. However, the evolutionary origin of domains themselves is more challenging and is largely unexplored area of research.

Domains often adopt to a limited number of structural forms called folds, despite the seemingly endless diversity of the proteins. These folds are largely formed by a limited 'vocabulary' of recurring supersecondary structural elements, often by repetition of the same element and, increasingly, elements similar in both structure and sequence are discovered. This suggests that modern protein domains evolved by fusion and recombination from a more ancient peptide world and that many of the core folds observed today may contain homologous building blocks.

Solenoid repeat proteins of Tetratrico Peptide Repeat (TPR) domain represent an attractive model to explore this issue. TPR domains are formed by repetition of an $\alpha\alpha$-hairpin, a supersecondary structural element. Since $\alpha\alpha$-hairpins are frequent in proteins, therefore TPR-like domains might have arisen by the repetition of protein fragments that were originally used in a different structural context.

In order to explore this question, we require a better ability to judge, which $\alpha\alpha$-hairpins are TPR-like. Currently, several resources are available for the prediction of TPRs, however, they often fail to detect divergent repeat units. We therefore developed 'TPRpred', a profile-based method which uses a P-value-dependent score offset to include divergent repeat units, and also exploits the tendency of the repeats to occur in tandem. We benchmarked the performance of TPRpred in detecting TPR-containing proteins and in delineating the individual repeats within a protein, against currently available resources. TPRpred not only performed significantly better in detecting divergent repeats in TPR-containing proteins, but also detected more number of individual repeat units.

We identified several promising $\alpha\alpha$-hairpins in non-TPR proteins which resemble the repeating unit of TPR, by using TPRpred in conjunction with structure-structure comparisons, and we further selected the best five hairpins namely, the mitochondrial outer membrane translocase Tom20, the ribosomal protein S20 (RPS20), the phospholipase C (PLC), the heat shock protein 20 (HSC) and the bacterial glucoamylase (BGA), to experimentally construct new TPR-like domains by repetition. Using each of these hairpins, we constructed three different artificial genes coding for one, two and three copies. The resulting artificial proteins were expressed, purified and then characterised using circular dichroism, thermal denaturation and fluorescence spectroscopy experiments. The biophysical properties of these TPR-like domains can also be correlated to the statistical significance of the parental hairpin likely to be a repeating unit of TPR. Although high-resolution structures have not yet been determined, proteins made from the hairpins of Tom20 and RPS20 appear to have native-like properties. The hairpin of RPS20 is significant in our study, because ribosomal proteins are among the most ancient proteins known, and since many of the modern non-ribosomal proteins contain fragments from the ribosomes, they might have been the building blocks in early protein domain evolution.

Zusammenfassun

Proteine stellen die am häufigsten vorkommende Gruppe der Biomoleküle, die aufgrund ihrer Diversität an der großen Mehrzahl der biochemischen Prozesse beteiligt ist. Die Faltungseinheit der Proteine ist die Domäne. Neuartige Proteine entstehen oft aus der Rekombination, dem Zufügen oder Entfernen vorhandener Domänen; sie sind daher stabile Bausteine der Evolution. Wie Domänen, die schon eine beträchtliche Komplexität haben, selbst entstanden sind, ist allerdings weitgehend unbekannt.

Die scheinbar endlose Vielfalt der Proteine reduziert sich auf eine begrenzte Zahl struktureller Formen, sogenannte Folds. Folds setzen sich aus Supersekundärstrukturen zusammen, die in einigen Fällen auch aus repetitiven Einheiten bestehen. Dies weist darauf hin, dass sie durch Fusion und Rekombination dieser Einheiten entstanden sein könnten.

Solenoidproteine, die aus sich wiederholenden Einheiten von Tetratricopeptiden (TPR) bestehen, stellen ein attraktives Modell dar um diese Frage zu untersuchen. TPR Domänen sind aus repetitiven $\alpha\alpha$-hairpins geformt, die als einzelne Elemente häufig in anderem Kontext in Proteinen vorkommen. Die Wiederholung und Verknüpfung von Proteinfragmenten, die ihren Ursprung in anderen Polypeptiden haben, könnte somit, nicht nur für TPR Domänen, ein wichtiges Prinzip der Evolution von Folds und Domänen darstellen.

Zur Beantwortung dieser Frage benötigen wir die Kenntnis, welche $\alpha\alpha$-hairpins TPR-ähnlich sind. Da die verfügbaren Resourcen oft divergierende Repeats nicht erkennen, haben wir “TPRpred” entwickelt, eine Methode auf der Basis von Profilen, die hierzu in der Lage ist.

TPRpred war nicht nur besser im Erkennen divergierender Repeats in TRP Proteinen, sondern erkannte auch eine höhere Zahl einzelner Repeat-Einheiten.

Wir identifizierten in nicht-TPR Proteinen mehrere $\alpha\alpha$-hairpins, die einer TPR Einheit ähnelten, und wählten für weitere Untersuchungen die besten fünf aus: Mitochondriale Auenmembrantranslokase Tom20, ribosomales Protein S20 (RPS20), Phospholipase C (PLC), Heat shock protein 20 (HSC) und bakterielle Gucoamylase (BGA). Mit diesen Hairpins konstruierten wir jeweils drei künstliche Gene mit einer, zwei bzw. drei verknüpften Einheiten. Die resultierenden Proteine wurden nach Expression in Escherichia coli gereinigt und biophysikalisch charakterisiert. Die Eigenschaften dieser TPR-ähnlichen Domänen korrelieren mit der statistischen Signifikanz, mit der sie der TPR-Einheit ähneln. Proteine, die aus Tom20 und RPS20 hervorgingen, haben vermutlich nativen Charakter, entsprechend einem gefalteten Protein. RPS20 ist auch deswegen bedeutsam, da ribosomale Proteine mit die ältesten bekannten Proteine sind, deren Fragmente daher die Bausteine in der frühen Evolution von Domänen gebildet haben könnten.

Contents

1 Introduction **1**
- 1.1 The origin of life 1
  - 1.1.1 The heterotrophic origin of life 2
  - 1.1.2 RNA world 3
  - 1.1.3 DNA-protein based life 5
- 1.2 Protein structure and folding 6
  - 1.2.1 Structural hierarchy 7
  - 1.2.2 Folding 8
    - 1.2.2.1 Correct folding 11
    - 1.2.2.2 Misfolding 11
- 1.3 Protein homology detection 13
  - 1.3.1 Finding similarities and inferring homologies 13
  - 1.3.2 Protein structure comparison 14
- 1.4 Protein engineering and design 15
- 1.5 Evolution 16
  - 1.5.1 Role of repetition in domain evolution 18
- 1.6 The solenoid proteins 19
- 1.7 Tetratrico Peptide Repeats (TPRs) 21
  - 1.7.1 TPR detection from protein sequences 21
  - 1.7.2 Origin of TPR domain 24
- 1.8 Aims of this study 24

2 Materials and Methods **26**
2.1 Computational procedure . . . 26
2.1.1 Algorithm . . . 26
2.1.2 Seed alignment construction . . . 27
2.1.3 Profile construction . . . 28
2.1.4 HMM Logos construction . . . 28
2.1.5 Searching for TPR-like hairpins . . . 28
2.1.5.1 Computing average TPR unit . . . 28
2.1.5.2 Searching for TPR-like hairpins . . . 29
2.1.6 Backtranslation . . . 29
2.1.7 Modelling . . . 29
2.2 Experimental procedure . . . 30
2.2.1 Plasmid construction . . . 30
2.2.1.1 Tom20-1, Tom20-2 and Tom20-3 genes . . . 30
2.2.1.2 RPS20-1, RPS20-2 and RPS20-3 genes . . . 31
2.2.1.3 PLC-1, PLC-2 and PLC-3 genes . . . 33
2.2.1.4 HSC-1, HSC-2 and HSC-3 genes . . . 34
2.2.1.5 BGA-1, BGA-2 and BGA-3 genes . . . 36
2.2.2 Transformation . . . 37
2.2.3 Plasmid DNA extraction . . . 38
2.2.4 DNA sequencing . . . 38
2.2.5 Protein production . . . 38
2.2.6 Protein concentration determination . . . 39
2.2.7 Tricine-urea-SDS-PAGE . . . 39
2.2.8 Protein purification . . . 40
2.2.8.1 Tom20-1, Tom20-2 and Tom20-3 proteins . . . 40
2.2.8.2 RPS20-1, RPS20-2 and RPS20-3 proteins . . . 41
2.2.8.3 BGA-1, BGA-2 and BGA-3 proteins . . . 41
2.2.8.4 Purification under denaturing conditions and refolding . . . 41
2.2.9 Circular Dichorism (CD) . . . 42
2.2.10 Fluorescence spectroscopy . . . 42
2.2.11 Chaperone Assays . . . 42

3 Results **43**
- 3.1 Bioinformatics analysis . . . 43
  - 3.1.1 Structure-based sequence alignment . . . 43
  - 3.1.2 Profile generation and selection . . . 44
  - 3.1.3 Analysis of the profiles . . . 49
  - 3.1.4 Benchmarking . . . 51
    - 3.1.4.1 Comparison of TPRpred and HMMER . . . 51
    - 3.1.4.2 Comparison of the web-servers using STAND family members . . . 53
    - 3.1.4.3 Comparison of the web-servers using known protein structures . . . 55
  - 3.1.5 TPR-like hairpins in globular proteins . . . 56
  - 3.1.6 Design of TPR-like domains . . . 61
    - 3.1.6.1 Design of Tom20-1, Tom20-2 and Tom20-3 . . . 66
    - 3.1.6.2 Design of RPS20-1, RPS20-2 and RPS20-3 . . . 69
    - 3.1.6.3 Design of PLC-1, PLC-2 and PLC-3 . . . 70
    - 3.1.6.4 Design of HSC-1, HSC-2 and HSC-3 . . . 75
    - 3.1.6.5 Design of BGA-1, BGA-2 and BGA-3 . . . 76
  - 3.1.7 Model structure of 3 repeat proteins . . . 79
- 3.2 Experimental analysis . . . 79
  - 3.2.1 Biochemical characterization . . . 84
    - 3.2.1.1 CD spectroscopy . . . 84
    - 3.2.1.2 Thermal denaturation . . . 91
    - 3.2.1.3 Trp fluorescence . . . 91
  - 3.2.2 In vitro chaperone activity of Tom20-1/2/3 . . . 94

4 Discussion **97**
- 4.1 The detection of TPRs . . . 97
- 4.2 Evolution of protein domains . . . 99

References **120**

Chapter 1

Introduction

1.1 The origin of life

It has been long recognized that deciphering the history of life on Earth is a profound task. The scientific field devoted to the origin of life on Earth is very young, having taken its first experimental steps in the 1950s. Though the question has captivated human imagination since the dawn of history, its scientific pursuit depended on several crucial conceptual developments during the 20th century. First, the emergence of life was conceived as an integral part of the general process of evolution, leading from the geochemistry of the barren Earth to the universal common ancestor, which later diversified into the Darwinian tree of life as illustrated in Figure 1.1. Following the rise of molecular biology in the 1950s and 1960s, the origin-of-life question could be formulated in biochemical and genetic terms, making it a subject of experimental investigation.

Early on, most scientists engaged in this research were chemists who attempted to formulate plausible scenarios for the prebiotic synthesis of organic building blocks, biologically relevant polymers, and the first metabolically or genetically functional chemical structures. In the late 1970s, however, geologists also became increasingly involved in the field. Their participation was associated with the rise of a new paradigm positing that the synthesis of organic building blocks and the emergence of life itself took place not in the ”primordial soup” of the traditional hypotheses but in the vicinity of undersea hydrothermal vents, at high temperature and under extreme pressure. Supporters of this new conception

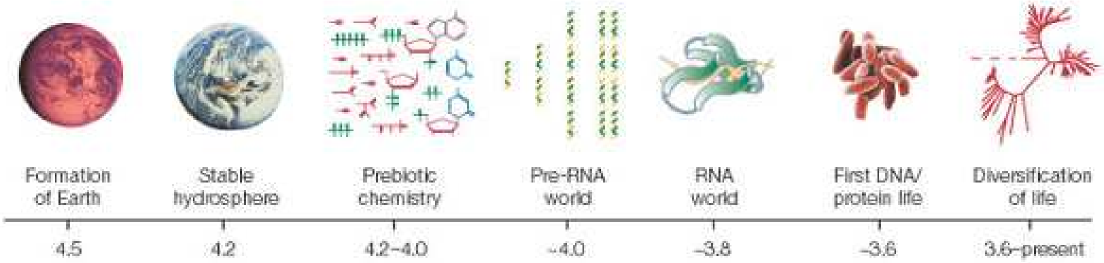


Figure 1.1: Timeline of events pertaining to the early history of life on Earth, with approximate dates in billions of years before the present [5].

claim that origin-of-life theories can now be subjected to more rigorous constraints posed by specific primordial physical settings [1]. On the other hand, the "soup people" in particular, Stanley Miller, renowned pioneer of the 1953 prebiotic simulation experiments, and his colleagues rejected the alternative paradigm as empirically untenable [2]. They argued that organic compounds are decomposed at 350°C rather than synthesized, and polymers such as peptide, RNA, and DNA are hydrolyzed rapidly rather than synthesized at vent temperatures [3]. This means that organic compounds would not accumulate over very long periods of time, and therefore the vent destruction sets a time frame for the origin of life of approximately ten million years.

We do not have a detailed knowledge of the processes that led to the appearance of life on Earth. However, there is convincing paleontological evidence showing that stromatolite-building phototactic prokaryotes were already in existence 3.5 billion years ago [4].

1.1.1 The heterotrophic origin of life

The Oparin-Haldane heterotrophic theory of the origin of life has been widely accepted on the basis that a heterotrophic organism is simpler than an autotrophic one, and prebiotic synthesis experiments show how easy it is under reducing conditions to produce organic compounds, many of which are used in present biology. However, there are, some recent examples of autotrophic proposals made for a variety of reasons.

One reason for proposing an autotrophic origin is the CO_2-rich model of the primitive Earth's atmosphere [6]. High pressures of CO_2 (10-100 atm) imply the absence of reducing conditions and organic compound synthesis, and therefore it would be necessary for the first organisms to biosynthesize their organic compounds, or to make use of the very small amounts of organic compounds brought in by comets and meteorites.

There have been so many unsuccessful attempts to produce prebiotic organic compounds with $CO_2+N_2+H_2O$ mixtures (in the absence of hydrogen) that one wonders whether successful prebiotic syntheses are possible under such conditions. The autotrophic theories are not supported by adequate experimental evidences of how organic compounds can be produced, and how such systems can work. This is quite a challenge, since even heterotrophic entities, which need to take their compounds from the environment, are difficult to envision.

1.1.2 RNA world

The discovery of catalytic RNA (ribozymes) gave credibility to prior suggestions that the first living organisms were self-replicating RNA molecules with catalytic activity, a situation called the "RNA world" [7]. It is unlikely that RNA itself with four bases (AUGC) and a ribose phosphate backbone was a prebiotic molecule (Figure 1.2). The period when the informational macromolecule had a backbone different from ribose phosphate and possibly different bases is refered as the "pre-RNA world" [8]. The "pre-RNA world" is assumed to have the same essential characteristic of the "RNA world" phenotype and genotype both reside in the same polymer, so no protein or related catalysts are required to be synthesized.

The "RNA world" idea has become widely accepted, because the RNA molecule has a pervasive role in contemporary biology, especially with regard to the most fundamental and highly conserved cellular processes. It is involved as a primer in DNA replication, a messenger that carries genetic information to the translation machinery, and a catalyst that lies at the heart of the ribosome. RNA instructs the processing of precursor messenger RNAs during splicing and editing, and mediates numerous other transactions of RNA and proteins in the cell [5]. It

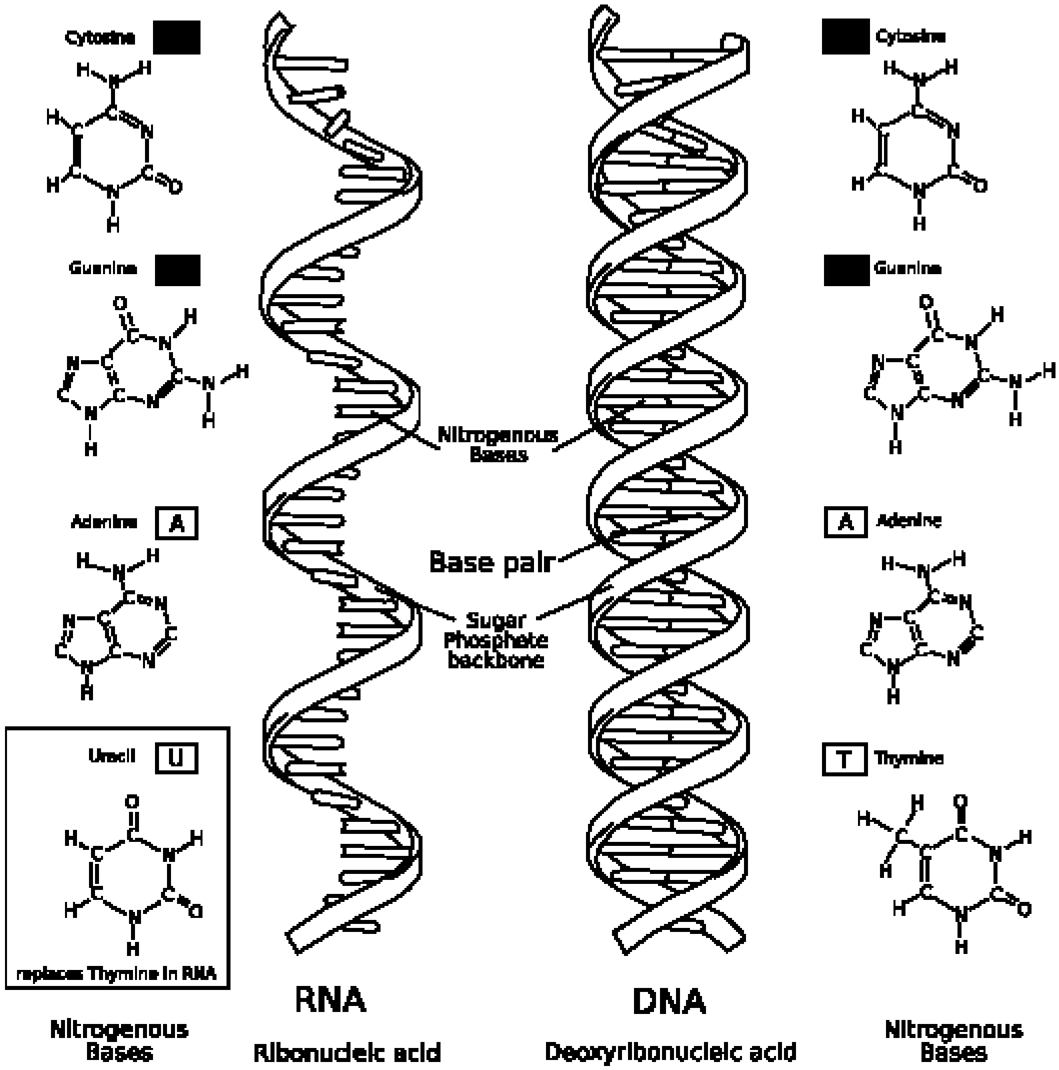


Figure 1.2: **The backbone structure of RNA and DNA. RNA with its nitrogenous bases to the left and DNA to the right.**

is as if a primitive civilization had existed prior to the start of recorded history, leaving its mark in the foundation of a modern civilization that followed. Although there may never be direct physical evidence of a RNA-based organism, because the “RNA world” is likely to have been extinct for almost four billion years, molecular archaeologists have uncovered artifacts of this ancestral era, none more pronounced than the recently reported crystal structure of the ribosome [9, 10, 11]. This structure reveals the face of the “RNA world” in the active role that RNA has in protein synthesis, and are considered as a relics or molecular fossils of the “RNA world” [12]. RNA folds into a variety of complex tertiary structures, analogous to structured proteins, and catalyses a broad range of chemical transformations [13]. It seems likely that RNA has the capability to support life based on RNA genomes that are copied and maintained through the catalytic function of RNA. However, there are substantial gaps in scientific understanding concerning how the “RNA world” arose, the degree of metabolic complexity that it attained, and the way that it led to DNA genomes and protein enzymes (Figure 1.1).

1.1.3 DNA-protein based life

Although RNA is well suited as a genetic molecule and can evolve to perform a broad range of catalytic tasks, it has limited chemical functionality and thus may not be equipped to meet certain challenges and opportunities that arise in the environment [14]. The crowning achievement of the “RNA world” was the invention of protein synthesis, instructed and catalysed by RNA, but also began its demise [15].

RNA is capable of performing all of the reactions of protein synthesis. The messenger, transfer and ribosomal RNA molecules that exist in all known organisms direct the assembly of specific polypeptide sequences, instructed by corresponding RNA sequences. The activation of amino acids in the form of aminoacyl adenylates, and subsequent transfer of the amino acids to the tRNAs, are catalysed in modern biology by the set of 20 aminoacyl-tRNA synthetase proteins. The final step of protein synthesis involves binding aminoacyl and peptidyl oligonucleotides at adjacent positions along the RNA template and catalysing peptide bond formation through attack of the α-amine of the amino acid on the carbonyl of the peptidyl ester [16].

The next step towards the origin of the genetic code was the formation of peptide bonds between amino acids that were attached to RNA. The products of this reaction must have conferred some selective advantage, even though the peptides probably would have been too small and too heterogeneous in sequence to function as catalysts [17]. Instead, they might have served as cofactors for ribozymes or been more effective than amino acids for any of the roles suggested above [18]. RNA-catalysed peptide bond formation might have resulted in a large number of possible peptide sequences, and even this mixture may have been useful [19]. However, the development of a crude mechanism for controlling the diversity of possible peptides would have been advantageous, and progressive refinement of that mechanism would have provided further selective advantage. It is reasonable to postulate that, like the modern translation apparatus, the ancestral translation system made use of messenger-like RNA molecules to gather aminoacyl-RNAs in a specific order through Watson-Crick pairing interactions. However, it is not clear how the detailed assignments of the genetic code were made [20].

It is not known whether the invention of protein synthesis preceded or followed the invention of DNA genomes. The primary advantage of DNA over RNA as a genetic material is the greater chemical stability of DNA, allowing much larger genomes based on DNA. Protein synthesis may require more genetic information than can be maintained by RNA.

1.2 Protein structure and folding

The name protein is derived from the Greek word “protas” which means ”of primary importance”. It is a complex high-molecular-mass organic compound that consists of amino acids arranged in a linear chain. Proteins were discovered by Jöns Jakob Berzelius in 1838 and are among the most actively-studied molecules in biochemistry [21]. Like other biological macromolecules such as polysaccharides, lipids, and nucleic acids, proteins are essential components of all living organisms [22]. Many proteins are enzymes that catalyze biochemical reactions, whereas many other proteins have structural or mechanical roles, such as those that comprise a cell's cytoskeleton (effectively a system of scaffolding that maintains the cell's shape and size). Proteins are also important components of cell signaling, the immune response, cell adhesion, the cell cycle, and essentially every process within a living cell.

1.2.1 Structural hierarchy

Proteins are enormously diverse. There are billions of species on Earth, and each codes for thousands of proteins. Although, these proteins superficially looks different, they often display substantial similarity both in sequence and three-dimensional structure [23]. All the proteins are linear heteropolymers built from 20 different L-α-amino acids [24]. Each of these amino acids share common structural features that includes an α carbon to which an amino group, a carboxyl group, and a variable side chain are bonded. Amino acids in a protein are linked by a peptide bond formed by a dehydration reaction. Once linked in the protein chain, an individual amino acid is often called a residue [25]. The peptide bond has two resonance forms that contribute to the double bond character and restricts rotation around its axis, forcing the alpha carbons to be roughly coplanar [26]. By convention the bond angles resulting from rotations at C_α are labeled ϕ (phi) for the N-C_α bond and ψ (psi) for the C_α-C bond. Both ϕ and ψ dihedral angles in the peptide bond are key determinants of the secondary structure assumed by the protein's backbone [27]. Due to the chemical structure of the individual amino acids, the overall polymer chain has directionality. The end of the protein with a free amino group is known as the N-terminus or amino terminus, while the end of the protein with a free carboxyl group is known as the C-terminus or carboxy terminus. The distinct structural hierarchy displayed in proteins is shown in Figure 1.3:

- **Primary structure**: the order in which the individual amino acids are arranged in a single protein molecule.
- **Secondary structure**: regularly repeating, non-cooperatively-folding, local structures stabilized by hydrogen bonds. The most common examples

are the α-helix and β-strand, though random coil regions with no defined hydrogen bonding pattern or characteristic shape are also frequent. Many different individual secondary structures can be present in the same protein molecule, because secondary structures are local.

- **Supersecondary structure**: sum of the combinations of secondary structures which are frequent. The most frequent supersecondary structures are $\alpha\alpha$-hairpins, $\beta\beta$-hairpins, and $\beta\alpha\beta$-elements.
- **Domain**: autonomously folding portion of a single protein molecule with the spatial relationship of the secondary structures to one another. Domain is generally stabilized by nonlocal interactions, most commonly the formation of a hydrophobic core, but also through salt bridges, hydrogen bonds, disulfide bonds, and even post-translational modifications. Many domains adopts limited number of structural form called "fold".
- **Tertiary structure**: the overall shape of a single protein molecule which constitutes the spatial relationship of the secondary structures and domains to one another. Tertiary structure is generally stabilized by nonlocal interactions.
- **Quaternary structure**: structure that results from the interaction of more than one protein molecule, usually called protein subunits, which function as part of the larger assembly or protein complex (oligomers). The complex made up of same subunits are called homomers, while with different subunits are called heteromers.

1.2.2 Folding

The covalent backbone of a typical protein contains thousands of individual bonds, where free rotation is possible around many of these bonds, therefore protein can assume an unlimited number of conformations. However, each protein has a specific function, strongly suggesting that each has a unique three-dimensional structure [28]. This is achieved by process called protein folding by which a protein structure assumes its functional shape or conformation [29] by

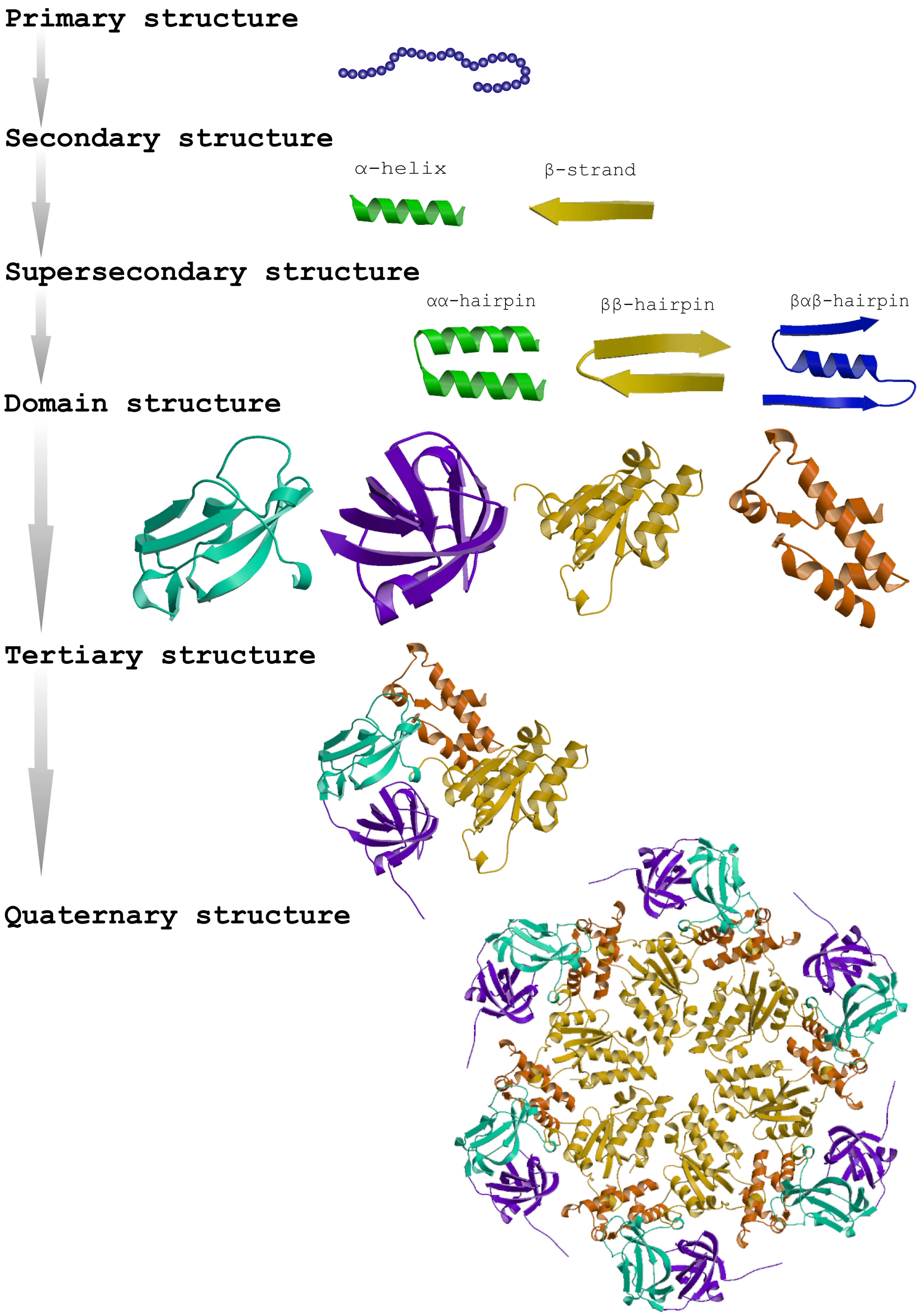


Figure 1.3: **Structural hierarchy in proteins.**

coiling and folding into a specific three-dimensional shape they are able to perform their biological function [30]. Protein folding is essential for life, yet the concentrated and complex interior of a cell is an inherently hostile environment for the efficient folding of many proteins. The reverse of this process is protein denaturation, whereby a native protein is caused to lose its functional conformation, and become a linear amino acid chain. Denatured proteins may lose their solubility, and precipitate, becoming insoluble solids. In some cases, denaturation is reversible, and proteins may refold, this process is called renaturation. In many other cases, however, denaturation is irreversible [31, 32].

The particular amino-acid sequence of a protein predisposes it to fold into its native conformation or conformations [33]. Many proteins do so spontaneously during or after their synthesis inside cells. While these macromolecules may be seen as folding themselves, in fact their folding depends a great deal on the characteristics of their surrounding solution, including the identity of the primary solvent (either water or lipid inside cells), the concentration of salts, the temperature, and molecular chaperones [34]. Some proteins constrained by sequence, topology, size, and function simply cannot fold by themselves and are instead prone to misfolding and aggregation. This problem is so deeply entrenched that a specialized family of proteins, known as molecular chaperones, evolved to assist in protein folding [35]. The large, oligomeric, and energy utilizing chaperonins or Hsp60s are one of the main essential class of molecular chaperones [36]. The bacterial chaperonin GroEL, along with its co-chaperonin GroES, is probably the best-studied example of this family of protein-folding machine [37].

Folding is a spontaneous process. The passage of the folded state is mainly guided by Van der Waals forces and entropic contributions to the Gibbs free energy (G): an increase in entropy is achieved by moving the hydrophobic parts of the protein inwards, and the hydrophilic ones outwards. This endows surrounding water molecules with more degrees of freedom. During the folding process, the number of hydrogen bonds does not change appreciably, because for every internal hydrogen bond in the protein, a hydrogen bond of the unfolded protein with the aqueous medium has to be broken.

Folding can funnel to a single stable state by multiple partially folded forms (molten globule) in conformational space [38, 39]. The general energy landscape picture provides a conceptual framework for understanding folding kinetics (Figure 1.4) [40, 41]. The predominant equilibrium in proteins is not between native and unfolded states, it is between the native and multiple partially folded forms. Some of these partially folded forms can be energetically close to the native state [42].

Results from many different disciplines including protein folding, NMR and fast kinetics are challenging the one sequence, one structure, one function paradigm. Proteins - the simplified view of which assumes a single rigid fold - have been shown to exist as an ensemble of different conformations in equilibrium before encountering a substrate. This has prompted a new view of protein structure and function in which conformational diversity provides a mechanism for controlling protein activation and permitting multi-functionality [43].

1.2.2.1 Correct folding

In certain solutions and under some conditions proteins will not fold at all. Temperatures above or below the range that cells tend to live in will cause proteins to unfold or ”denature” (this is why boiling makes the white of an egg opaque). High concentrations of solutes and extremes of pH can do the same. A fully denatured protein lacks both tertiary and secondary structure, and exists as a so-called random coil. Cells sometimes protect their proteins against the denaturing influence of heat with enzymes known as chaperones or heat shock proteins, which assist other proteins both in folding and in remaining folded. Some proteins never fold in cells at all except with the assistance of chaperone molecules, that either isolate individual proteins so that their folding is not interrupted by interactions with other proteins or help to unfold misfolded proteins, giving them a second chance to refold properly

1.2.2.2 Misfolding

Protein misfolding is now recognized to be a major contributing factor in a number of protein folding diseases, including amyotropic laterial sclerosis, cystic fibrosis, Alzheimer’s disease, Parkinson’s disease, and a host of many different amyloidosis diseases. These diseases are associated with the aggregation of misfolded

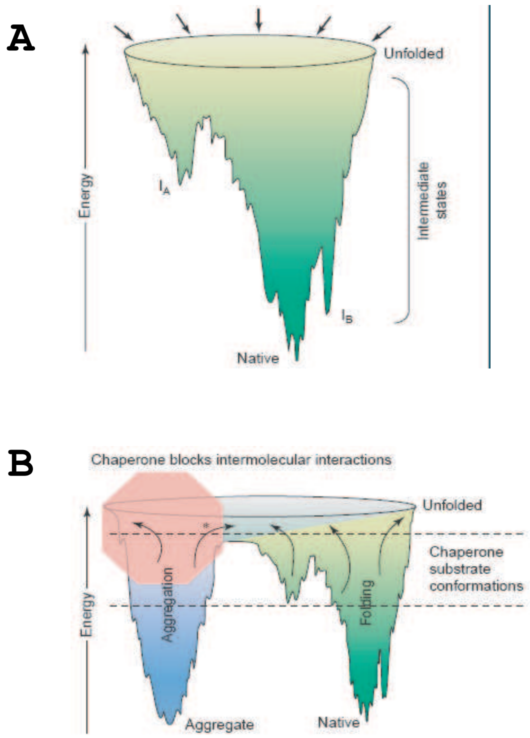


Figure 1.4: **Folding funnels to describe protein folding.** (A) A typical folding funnel diagram used to describe the folding of a well behaved (often single-domain) protein *in vitro*. (B) The effect of molecular chaperones on chain conformations in the context of a multiple-minima folding funnel. For simplicity, funnel space is shown only for stabilizing conformations. Hsp60 molecular chaperones recognize polypeptide chains (in either funnel) with conformations that expose significant amounts of hydrophobic surface area; these conformations are indicated by the area of the funnels within the broken lines. In the iterative annealing mechanism, the ATP hydrolysis cycle releases polypeptide chains with fewer stabilizing interactions or higher chain entropy (i.e. a higher position on the funnel surface, indicated by the curved black arrows), permitting a new path down the funnel surface that possibly traverses the barrier separating the folding and aggregation funnels (arrow marked with asterisk). In the Anfinsen cage mechanism [44], chain isolation in the central cavity of the chaperone effectively blocks the aggregation funnel (i.e. the conformations stabilized by intermolecular interactions, indicated by the pink octagon); this portion of the energy landscape becomes inaccessible to the chain. Figure adopted from Clark [45].

proteins into insoluble plaques; it is not known whether the plaques are the cause or merely a symptom of illness. Protein aggregation and misfolding reactions are also the bane of protein production and impede pharmaceutical drug development. Understanding the fundamental in vivo factors that control the kinetics of protein misfolding is the crucial aspect involved in developing procedures and strategies to avoid this deleterious side reaction. Elegant in vivo work of Nollen et al.[1] has shown that the elements controlling protein homeostasis such as protein synthesis, energy production in the cell, chaperone-dependent protein folding, protein transport, and protein degradation collectively control intracellular protein aggregation. However, it is also known that protein folding is influenced by the presence of intracellular osmolytes that, in turn, dramatically affect protein stabilities, protein folding rates, and protein aggregation.

1.3 Protein homology detection

Protein homology detection and sequence alignment are the basis of protein structure prediction, function prediction and evolution. Homologous proteins are those that are evolutionarily related. They usually perform the same function in different organisms. Homologous proteins from different organisms may have nearly similar amino acid sequences. Many positions in the amino acid sequence are occupied by the similar residues in all organisms are called conserved residues whereas the positions occupied with variable residues are called non-conserved residues.

1.3.1 Finding similarities and inferring homologies

Sequence similarity searching is an essential tool for molecular biologists to understand the function of yet uncharacterised gene [46]. Searches for homologous relationships based on sequence similarity have become a routine step. The mathematics behind sequence comparison has been discovered many times in different areas of science. Geology, bird and whale song analysis, and voice recognition are few such areas. The first description of a sequence similarity search method in biology was published by Needleman and Wunsch [47]. Traditional pairwise sequence alignment methods can be used to detect homology to sequences with obvious evolutionary relationships to a known structure. Generally, for sequences with identities > 30 %, fast sequence searching methods such as FASTA [48], and BLAST [49] are fairly capable at detecting related proteins by scoring pairwise comparisons and compare in accuracy to the slower, Smith and Waterman [50] based method SSEARCH [48]. However when sequence identities fall below 30 %, conventional pairwise sequence comparison methods fail to detect relationships [51], therefore, accurately annotating genes that encode proteins with low sequence identity to any known protein structure remains problematic.

Sequence searching has been improved beyond pairwise comparisons with the introduction of profile-sequence based methods such as PSIBLAST [52], ISS [53], SAM-T98 [54] and FFAS [55]. These methods use information from profiles of related sequences in order to detect more distant relationships. A significant improvement over profile-sequence based methods are made possible by comparing profiles to profiles. Several programs for homology recognition have been developed that are based on profile-profile comparison: LAMA [56] PROF_SIM [57], COMPASS [58] and HHsearch [59]. These programs are shown to be significantly more sensitive than PSI-BLAST and have been applied for identifying evolutionary links between protein families previously thought to be unrelated. Proteins may remain structurally very similar long after their sequence similarity has disappeared because structures diverge much more slowly than sequences [60].

1.3.2 Protein structure comparison

Proteins have complex three-dimensional shapes that, by eye, often bear striking similarity to one another over their entire lengths or over shorter regions. In parallel to what can be deduced from pure sequence relationships, structural similarities also suggest the possibility of evolutionary relationships between proteins, because it is widely accepted that structure is better conserved than sequence (at least given our current ability to detect sequence relationships). However, detecting geometric relationships between proteins is a far more uncertain process than the identification of pure sequence relationships, as the latter can be clearly defined in statistical terms. In contrast, there is considerable ambiguity in how to describe a geometric relationship between two proteins, resulting in the large number of approaches to this problem described in the literature [61].

Computer analysis of tertiary structures began with the systematic comparison of tertiary structures of oxy- and deoxyhemoglobin by least squares fitting of their electron densities [62]. Subsequently, comparison of protein tertiary structures was initiated as the results of X-ray analysis became available [63, 64, 65]. These methods may identify structural equivalence, which defines those structural elements that are coincident in three dimensions to generate structural alignments. Structural alignment programs define scoring functions that measure the geometric similarity between proteins and use various algorithms to search for two substructures such that these functions are optimal. Most existing similarity measures can be classified into two main types depending on what they compare: the distances between corresponding pairs of atoms in the two structures (e.g. DALI [66], CE [67] and SSAP [68]); and the relative positions of the corresponding atoms of two proteins that have been superimposed (e.g. PrISM [69] and SSM [70]). For evolutionary comparisons, a sequential series of structurally equivalent residues, is required where the chain segments are similarly directed. This is known as topological equivalence.

1.4 Protein engineering and design

Protein design is the design of new protein molecules from scratch. Protein engineering deals with the process of developing useful and valuable proteins. It is a young discipline, with much research currently taking place into the understanding of protein folding and protein recognition for protein design principles. The number of possible amino acid sequences is infinite, but only a subset of these sequences will fold reliably and quickly to a single native state. Protein design involves identifying such sequences, in particular those with a physiologically active native state. A major challenge in protein design is to create sequences that can fold uniquely, i.e. to a single conformation rather than to many conformations.

The 20 amino acids are enough to construct an astonishing number of different protein structures that carry out a staggering array of biochemical processes. The so-called protein-folding problem that has preoccupied structural biologists for more than four decades can be most simply explained as the problem of discovering how simple strings of 20 amino acids can encode the complex three-dimensional folded structures of proteins. Solving the protein-folding problem implies that we would have the ability to read any amino acid sequence and deduce the correct native folded structure for that protein. Even with recent progress [71], a complete solution to this problem lies somewhere in the future. Despite the difficulty in going directly from sequence to structure, a number of researchers have been interested in the inverse problem [72, 73, 74]: deducing an amino acid sequence that will, when synthesized, self-assemble into a single desired 3D structure. Unlike the protein-folding problem, which has only one desired solution (the native folded state of the protein), the inverse case is likely to have many solutions. There are many examples of pairs of proteins that fold in a strikingly similar way, but which have no evident similarity in their amino acid sequences [75]. Statistically at least, solving the inverse protein-folding problem with its many potential solutions should be easier than solving the protein-folding problem, which has usually just a single correct solution [76].

Protein designing greatly enhances our understanding of protein evolution by designing the novel protein molecules which were presumed to be existed during the evolution of protein of interest. A fascinating example is provided by the α/β-barrel enzymes HisA and HisF, implicated in the histidine biosynthesis in *Thermotoga maritima*. Despite a sequence identity of only 25 %, specific structural features strongly suggest that HisA and HisF have evolved from a common half-barrel precursor [77, 78].

1.5 Evolution

Complex organisms have evolved from a limited number of primordial genes and proteins [79]. However, the mechanisms by which the earliest proteins evolved and their role for the present diversity of protein function are still unknown. Protein evolution is the object of intense study. The reason for this interest lies in the inference of homology to explore life, based on the study of model systems. Particularly in molecular biology where searches for homologous relationships based on sequence similarity have become a routine step to understand the function of yet uncharacterised gene [46].

Proteins tend to change in the successive generations through basic mechanisms that produce evolutionary changes like natural selection (which includes ecological, sexual, and kin selection) and genetic drift; these two mechanisms act on the genetic variation created by mutation, genetic recombination and gene flow. Natural selection is the process by which individual organisms with favorable traits are more likely to survive and reproduce. If those traits are heritable, they pass them to their offspring, with the result that beneficial heritable traits become more common in the next generation [80]. Most frequently, changes result from point mutations, insertions and deletions. By these processes, proteins may become so dissimilar that their common origin cannot be detected from their sequences, even though they may still fulfill fundamentally the same function [81]. However, their structures diverge much more slowly, providing evidence of common ancestry long after their sequence similarity has decayed [60].

The protein complement of an organism is the result of parental inheritance, acquisition (through lateral transfer, viruses, or mobile elements) and duplication. Duplication is central to the diversification of proteins [82]. At the level of full genomes, duplication is an effective path to increased complexity, which repeatedly occured in the course of evolution [83, 84, 85]. At the level of operons, duplication may lead to the efficient evolution of novel pathways [86]. At the level of single genes, duplication allows the emergence of systems with complex functionality, such as the vertebrate olfactory system, which is built on thousands of homologous G-protein-coupled receptors. In each of these cases, the duplicated copies are freed from the selective pressure to maintain function and in fact come under pressure to assume a novel selectable function in order to avoid extinction through mutational inactivation [87].

Duplication, accompanied by gene fusion, is also essential for a variety of other processes that result in the generation of novel proteins, such as unequal recombination [88], circular permutation [89, 90], and domain shuffling [91]. Unequal recombination is the primary mechanism that gives rise to repetitive proteins [88, 92]; an extreme case is the giant muscle protein, titin, which consists of hundreds of immunoglobulin domains.

1.5.1 Role of repetition in domain evolution

Most of the protein parts are derived from a basic complement of autonomously folding units called domains. Protein domains are the basis for the classification of proteins into a hierarchy of families, superfamilies, and folds. The basic complement of domains was already established to a large extent at the time of the last common ancestor [93], but some very successful domains arose later within the bacteria, archaea, or eukaryotes and radiated into the other kingdoms by endosymbiosis or lateral transfer [79].

The basic processes of mutation, duplication, and shuffling have led from a set of ancestral domains to the complex proteins observed today. The ability of repetition to generate novel structures has been documented in only few cases [94,95]. Many of the most important fibrous proteins contain short peptide repeats(Figure 1.5b): (1) Collagen is formed by hundreds of proline- and hydroxyproline-rich Gly-X-Y repeats. (2) Coiled coils are primarily built on a repetitive pattern of seven residues in which the first and fourth residues are hydrophobic and the others are hydrophilic. (3) Finally, parallel β-helices are formed by stacked coils of two or three β-strands. The evolutionary success of repetitive proteins results from the fact that repetition intrinsically promotes stability through the periodic recurrence of favorable interactions [96].

Greater structural variability can be obtained by the repetition of larger units with defined secondary structure. Frequently, the most common units correspond to the same three supersecondary structures (Figure 1.5a). At the simplest level, the monotonous repetition of one supersecondary structure element generally gives rise to open-ended, solenoid structures (Figure 1.5c): Tetratrico Peptide Repeat (TPR), HEAT, Armadillo-, and Ankyrin-repeat proteins are formed of stacked $\alpha\alpha$-hairpins, leucine-rich repeat proteins are formed of $\beta\alpha\beta$ elements, and bacterial choline-binding domains are formed of $\beta\beta$-hairpins.

In some cases, repetition of these elements may lead to closed, globular structures (Figure 1.5d-g): for example TIM barrels ($\beta\alpha\beta$), and β-propellers ($\beta\beta$), both of which have yielded useful scaffolds for the emergence of catalytic activity and thus help trace a path of increased complexity from repetitive proteins to fully differentiated enzymes. Repetition is not only an important mechanism in the evolution of multidomain proteins, but also in the evolution of the domains themselves. The same repetitiveness is detectable in a large class of membrane-embedded domains, the porin β-barrels of bacteria and organelles (Figure 1.5g). These are formed of between four and eleven $\beta\beta$-hairpins in a circular arrangement [79].

†**Legend for Figure 1.5.** The panels show A: the three most important supersecondary structures; B: the main fibrous proteins (left-handed β-helix, 1L0S; right-handed β-roll, 1SAT; coiled coil, 1ZIK; collagen, 1BKV); C: solenoid proteins formed by repetition of supersecondary structure elements (stacked β-hairpin, 1HCX; TPR repeat, 1ELR; leucine-rich repeat, 1A4Y); D: superfolds with recognizable internal symmetry; the repeat unit is colored (β-trefoil, 4FGF; jelly-roll, 1GOH; immunoglobulin-like, 1JP5; TIM barrel, 1HTI; ferredoxin-like, 1APS; up-and-down four-helix bundle, 1RPR); E: superfolds without recognizable internal symmetry; some supersecondary structures are colored for illustration (OB-fold, 1QVC; UB-roll, 1LKK; globin, 1EBC; doubly wound, 5CHY); F: a β-propeller (1TBG); and G: the two types of membrane proteins: all-β (porin, 2POR) and all-α (rhodopsin, 1L9H).

1.6 The solenoid proteins

Solenoid repeat proteins have recently attracted interest because of their versatility as scaffolds for the engineering of protein-protein interactions [97]. This class of proteins is characterized by homologous, repeating structural units, which stack together to form an open-ended superhelical structure (Figure 1.6). Such an arrangement is in contrast to the structure of most proteins, which fold into a compact shape [98]. Solenoid structures adopt a variety of shapes, depending on the structural features of the repeating structural unit and the arrangement of individual units in the solenoid. The curvature created by the superhelical nature of these proteins predetermines the target proteins that can bind to them [99].

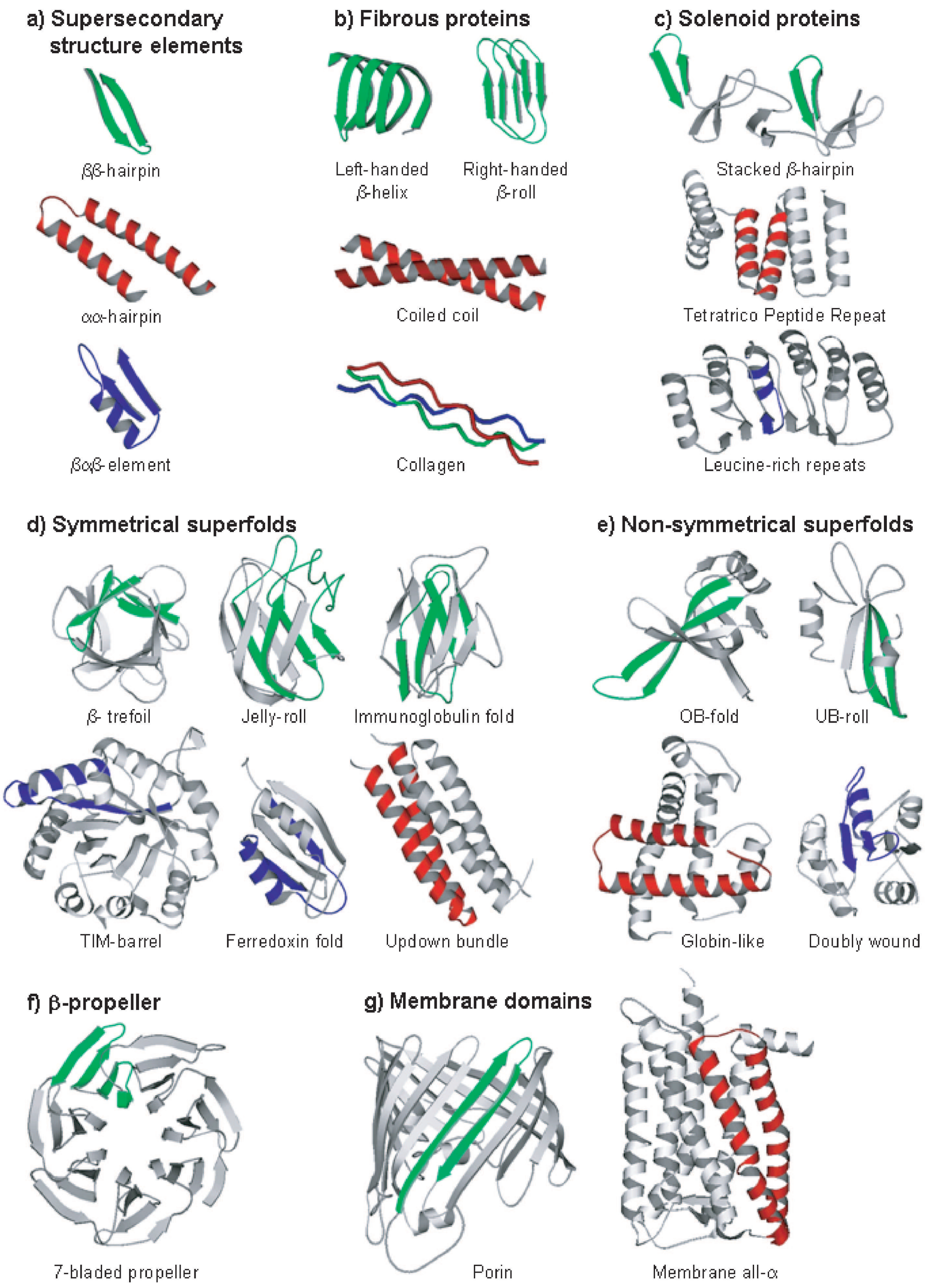


Figure 1.5: **Proteins from pieces.** Figure adopted from Lupas *et al.* [79]. †Legend of this Figure is shown in the previous page.

1.7 Tetratrico Peptide Repeats (TPRs)

Solenoid repeat proteins of the Tetratrico Peptide Repeat (TPR) family are involved as scaffolds in a broad range of protein-protein interactions [97, 100]. Sequence analysis of TPR-containing protein Cdc23p from *S.cervisiae* by Sikorski *et al.*, revealed a repeating 34 amino acid motif, which they named "Tetratrico Peptide Repeat" for the 34 residues constituting the repeating unit [101]. TPR-containing proteins have since been found in almost all organisms, ranging from bacteria to archaea and eukaryotes [102]. However, they are more abundant in eukaryotes. TPRs occur in a number of functionally diverse proteins, that are involved in a variety of biological processes like cell cycle control, transcription, protein import, protein folding, signal transduction and neurogenesis [103, 104].

The first crystal structure of the TPR-containing domain of human protein phosphatase 5 [105] revealed that, it is formed of one or more stacked helix-turn-helix ($\alpha\alpha - hairpin$) repeat units (Figure 1.6A). The repeat units are arranged in such a way that the polypeptide chain forms a continuous right-handed super-helical architecture. This architecture results in a continuous helical groove, that acts as a scaffold in mediating protein-protein interactions with the target proteins [106]. Moreover, this super-helical arrangement of a repeating structural unit is typical to all the members of α-solenoid proteins [99]. In natural proteins, the number of repeating units varies from 1-30. Comparison of TPRs from a variety of proteins reveals a high degree of sequence diversity with a conservation mainly in terms of size and hydrophobicity of a few key residues [102, 107].

1.7.1 TPR detection from protein sequences

The Tetratrico Peptide Repeats (TPRs) together with their related repeats, the Pentatrico Peptide Repeats (PPRs) and the SEL1-like repeats, form a large family within the solenoid repeat proteins. The repeating unit of TPRs, PPRs and SEL1-like repeats are formed of two or more stacked 34, 35 and 36-amino acid $\alpha\alpha$-hairpin repeat units, respectively [102, 108, 109]. These solenoid repeat proteins are involved in a diverse spectrum of cellular functions such as cell cycle control, transcription, splicing, protein import, regulatory phosphate turnover and protein folding, by virtue of their tendency to bind target proteins [110, 108, 111].

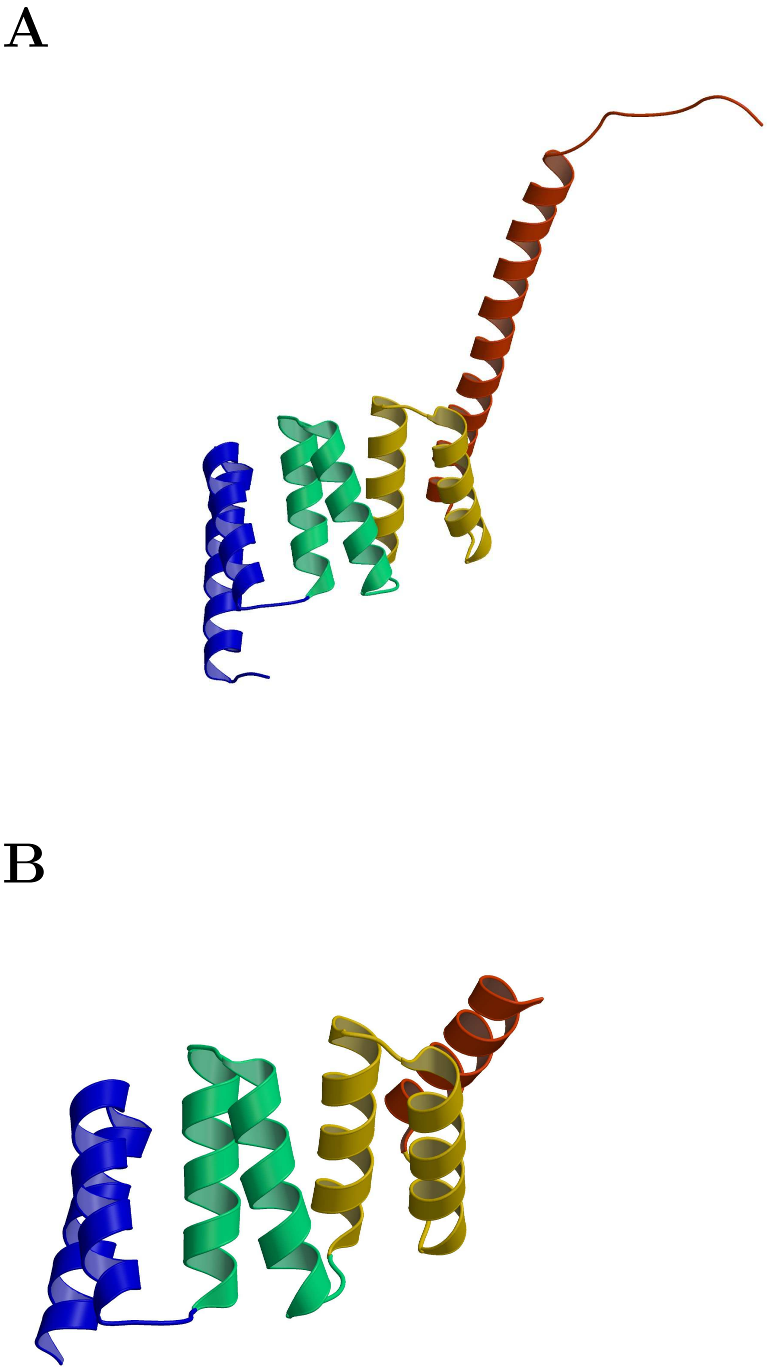


Figure 1.6: **Structure of TPR domains.** The individual repeat units are shown in different color. The extra helix present at the C-terminus is shown in red. (A) The three TPRs of protein phosphatase 5 (PDB code 1A17). (B) The three TPRs of consensus designed TPR domain (1NA0).

Homologous structural repeat units are often highly divergent at the sequence level, a feature that makes their prediction challenging. Currently, several web-based resources are available for the detection of TPRs, including Pfam [112], SMART [113], and REP [114]. These resources use hidden Markov model (HMM) profiles or sequence profiles, which are constructed from the repeats trusted to belong to the family. However, the profiles used are constructed from closely homologous repeats; therefore, divergent repeat units often get a negative score and are not considered in computing the overall statistical significance, even though they are individually significant. For this reason Pfam, SMART, and REP perform with limited accuracy in detecting remote homologs of known TPR-containing proteins and in delineating the individual repeats within a protein [115, 116]. For example, TPR-like repeats from the central domain of MalT protein [*E. coli*;PDB:1HZ4] are not detected by these resources. MalT is the transcription regulator of the maltose regulon, which is responsible for the uptake and catabolism of malto-oligosaccharides in *E. coli* [117].

In order to predict such divergent repeats, we have developed a specialized tool (TPRpred), which is able to predict TPR, PPR and SEL1-like repeats from protein sequences. The advantages of our method are the following:

- We construct optimized profiles through iterative searches by varying the threshold for inclusion of repeats into the profiles.

- We apply a score offset in such a way that repeats with P-value ≤ 0.01 will get a positive score. Therefore, even marginally significant repeats will contribute to the whole-protein P-value.

- Putative repeat units located near an already identified repeat get a tight-fit reward in order to account for the tendency of repeats to occur in tandem.

- Our tool reports not only P-values, based on the score distribution of true negatives derived from the known protein structures, but also computes a probability that a target sequence is a TPR protein.

1.7.2 Origin of TPR domain

Almost all known TPR-containing proteins can be detected using a single sequence profile containing positional residue conservation information from the homologous repeat units [118]. This suggests that, the common ancestor is likely to be a single repeat protein [119], from which multi-repeat proteins are thought to have evolved by gene duplication and fusion events [119]. These genetic events are crucial mechanisms of protein evolution, that multiply and link functional protein fragments [87, 120]. The other plausible evolutionary scenario is thought to be duplication and recombination of the α-hairpin protein fragments that were originally present in a different structural context [88, 121, 122, 79]. Although the repeat units are highly similar at structural level, no structural repeat unit is known that consists of identical sequence repeat units in naturally occurring TPR-containing proteins. However, recently perfectly repetitive TPR domains were engineered by consensus protein design that are identical both at the sequence and structural level (Figure 1.6B) [123, 124, 125]. We undertook a similar strategy but from an evolutionary perspective.

To address the evolution of TPRs, we have identified several α-hairpins by structure and sequence comparisons of known protein structures, which resemble the repeating unit of TPRs both at the structural and sequence level. We have selected the best five hairpins, which includes: mitochondrial outer membrane translocase Tom20 protein [102, 126], ribosomal protein S20 (RPS20), the phospholipase C (PLC), the heat shock protein 20 (HSC) and the bacterial glucoamylase (BGA). From each of these five hairpins, three separate proteins were constructed with one, two, and three copies of the respective hairpins.

1.8 Aims of this study

To validate the hypothesis that modern protein domains have evolved from a limited set of ancient peptides by fusion and recombination, using TPR domain as a case study. TPR domain is formed by the repetition of an $\alpha\alpha$-hairpin, a supersecondary structural element. Given that $\alpha\alpha$-hairpins are frequent in proteins, TPR-like domains might have arisen from the repetition of protein fragments that were originally used in a different structural context.

- In order to explore this question, we require a better ability to judge, which $\alpha\alpha$-hairpins are TPR-like. Currently, several web-based resources are available for the detection of TPRs, including Pfam, SMART, and REP. However, these bioinformatics tools perform with limited accuracy in detecting remote homologs of known TPR-containing proteins and in delineating the individual repeats within a protein. We therefore developed TPRpred, a profile-based method that uses a P-value-dependent score offset to include divergent repeat units, and exploits the tendency of repeats to occur in tandem. With this profile-to-sequence method we are able to identify members of the TPR family much more broadly.
- We attempted to build new TPR-like domains by repetition of $\alpha\alpha$-hairpins that resemble the repeating unit of TPR, from the best five identified hairpins of non-TPR proteins, obtained using TPRpred in conjunction with structure-structure comparisons. One of the significant hairpin is from the ribosomal protein S20 (RPS20), which is of particular interest since ribosomal proteins are the molecular fossils of the ancient peptide world.

Chapter 2

Materials and Methods

2.1 Computational procedure

2.1.1 Algorithm

Given a query sequence of length L and a sequence profile of length W representing a single repeat unit, TPRpred finds the best-scoring alignment of the sequence with an integer number of repeats, each of them aligned without internal gaps using standard log-odds scoring. Tandem repeats with a gap of $\leq K$ residues are rewarded by r bits, while gaps of $> K$ residues are not penalized ($K = 10$ in our benchmarks).

Since no internal gaps are allowed within repeats, the score distribution of the repeat profile with unrelated sequences has an almost perfect Gaussian distribution[1]. The λ and μ parameters of this distribution are derived from a calibration search against a database of unrelated protein sequences from the SCOP database [127]. The tails of a Gaussian distribution approach zero much faster than the tails of a Gumbel distribution (which would be obtained if internal gaps were allowed). Therefore, the same positive score of a true repeat unit will generally have a much higher significance in the case of a Gaussian as compared to a Gumbel distribution. Hence, the restriction of ungapped repeats increases the sensitivity of TPRpred for detecting ungapped repeat families such as TPRs,

---

[1] The score is a sum of W independent random variables and therefore it approaches a Gaussian according to the central limit theorem.

PPRs, SEL1-like and others with duplicated helical hairpins. TPRpred is able to calculate more realistic (i.e. less optimistic) E-values, by calibrating with true negative sequences as opposed to random sequences.

If the reward r for closely spaced repeat units is set low (e.g. zero) then one will fail to detect many repeats if their score is below zero. This is the case for the HMMER software [128], where often repeat instances have scores below zero even though their P-values are significant (e.g. below 0.01). Since alignment algorithms find the alignment with maximum score, they will skip repeat instances that are assigned negative scores. On the other hand, if r is set high, many false positive repeat units will be found within K residues of an already ascertained repeat unit. We therefore set the reward r such that the probability of finding a false positive repeat instance within K residues of another repeat is 0.01. To further increase sensitivity, we add an offset to the repeat unit match score in such a way that the probability for the observation of a repeat in an unrelated database protein is equal to 0.01. This ensures that even repeat units with no neighbours within K residues will get detected, if their P-value is better than 0.01, independent of the original score baseline (which depends on a null model that is not appropriate for this purpose). At the same time, this global offset guarantees that only very rarely (with probability $\approx 10^{-4}$) TPRpred will find more than one false positive repeat unit in an unrelated protein. TPRpred not only computes P-values, which are solely based on the true negative score distribution, but is also able to report the probability that a target sequence is a true homolog, by making use of both the true positive and true negative score distributions.

The algorithm has been implemented as a computer program "TPRpred", written in C++ (a Perl version is also available) and has been tested on a GNU/Linux platform with a i386 processor architecture.

2.1.2 Seed alignment construction

We obtained all those three-dimensional structures that are classified as TPRs family (SCOP:a.118.8.8) in the Structural Classification of Proteins (SCOP) database version 1.59 [127], from the Protein Data Bank (PDB) [129]. Each of the TPR-containing domains were further divided into individual repeat units, and then superimposed interactively in Swiss-PDB Viewer [130] using the first repeat of protein phosphatase 5 (1A17, chain A) from *Homo sapiens* as a reference structure. Structure-based sequence alignment was generated. The same procedure was applied to obtain structure-based sequence alignment for SEL1-like repeat family. For the PPR family the seed alignments from the Pfam was used [112].

2.1.3 Profile construction

From each alignment, profile was built by using the Ppmake program of TPRpred suite. The repeat units with 70 % maximum sequence identity were included in profile construction to ensure that the profile did not include very similar repeats. To avoid overfitting and to make the profile more generalised Ppmake takes into consideration the pseudocount and sequence weighting using the Gonnet substitution matrix with 17 % underlying sequence identity.

2.1.4 HMM Logos construction

Multiple sequence alignments correspond to best profiles of TPRs, PPRs and SEL1-like repeat families were used to create profile HMMs using HMMER suite version 2.3.2 [128]. Resultant profile HMMs were submitted to HMM Logos webserver with default options to draw HMM Logos [131].

2.1.5 Searching for TPR-like hairpins

2.1.5.1 Computing average TPR unit

All the three-dimensional (3D) structures were obtained from the Protein Data Bank (PDB) [129] which belong to the TPR-like superfamily (SCOP:a.118.8) in the Structural Classification of Proteins (SCOP) database (version 1.65) [127]. Each of the TPR-containing domains were further separated into individual repeat units. The average coordinates for the TPR unit were derived from the superposition of these repeat units using the STAMP program [132].

2.1.5.2 Searching for TPR-like hairpins

The coordinate files for the structures classified in SCOP (version 1.65) were obtained from the PDB by excluding all the TPRs. The C_β atom coordinates were extracted into a new database using an in-house Perl script. Similarly, the C_β atom coordinates were also extracted from the average TPR unit coordinate file. Subsequently, this was used as a query to search the database by employing the contact map description. The overlap of two contact maps, as defined by the number of contacts between equivalenced residues in two proteins that are simultaneously present in both structures was used as a measure of similarity between two protein structures, as described previously by Godzik *et al.*, [133]. The computational algorithm was implemented in an in-house C++ program.

2.1.6 Backtranslation

The amino acid sequence which corresponds to TPR-like repeating units were reverse translated into nucleotide sequence using an in-house Perl script. The perl script uses the codon usage table of *Escherichia coli K12* available from the Japanese codon usage database (http://www.kazusa.or.jp/codon/) [134]. The script read in the format of table that is compatible with that of CodonFrequency output in the GCG Wisconsin Package. This script also has an option to assign the different synonymous codon for the given protein sequence.

2.1.7 Modelling

The target protein sequences were aligned to designed TPR protein (1NA0). In case of Tom20 and RPS20 proteins 1OM2 and 1FJGT structure also used in generating target-template alignments. The refined sequence-structure alignment was used by MODELLER [135] to construct a 3D model of the target protein. Model building began by extracting distance and dihedral angle restraints on the target sequence from its alignment with the template structure. These template-derived restraints were combined with most of the CHARMM energy terms to obtain a full objective function. Finally, this function was optimized by conjugate gradients and molecular dynamics with simulated annealing to construct a model that satisfied all the spatial restraints.

2.2 Experimental procedure

2.2.1 Plasmid construction

We obtained restriction and DNA modifying enzymes from New England Biolabs, Pfu DNA polymerase from Finnzymes, pET-28b vector from Novagen. DNA manipulations were performed using standard molecular techniques. The genes encoding the proteins were produced by polymerase chain reaction (PCR) using corresponding primers with appropriate restriction sites at their termini to allow ligation into an N-terminal 6×-his-tagged pET28b vector (Novagen), that is under the control of T7 promoter. All the synthetic oligonucleotide sequences are given in standard $5'$ to $3'$ direction. The details of individual plasmid construction is described below.

2.2.1.1 Tom20-1, Tom20-2 and Tom20-3 genes

Expression plasmids encoding one, two, and three repeat genes were constructed as follows using the primers mentioned in Table 2.1. (A) The gene encoding the single repeat (Tom20-1) was amplified by PCR using cDNA from brain tissue of *Rattus norvegicus* as a template with Tom20-1_fr and Tom20-1-stop_rv primers. The PCR product was purified by agarose gel electrophoresis followed by Quiaquick column purification from agarose gel slices (Gel extraction kit, Quiagen). This product was double digested with NdeI/XhoI and following purification with Quiaquick purification columns (Quiagen) the DNA was ligated to NdeI/XhoI-digested pET28b vector (Invitrogen). Thereby Tom20-1 encoding gene was cloned in frame with the 6×-His tags of the vector plasmid. (B) A portion of the gene (one of two repeats) encoding two repeat (Tom20-2) was amplified by PCR using pET28b-containing Tom20-1 gene as a template with Tom20-1-cap_fr and Tom20-1-stop_rv primers. The other portion of the gene (two of two repeats) was produced by PCR with Tom20-2_fr and Tom20-2_rv overlapping primers. Finally, the full gene encoding Tom20-2 gene was produced by an additional PCR reaction with Tom20-2_fr and Tom20-1-stop_rv primers corresponding to the gene termini, and by using the previous two independent PCR reaction products as templates. The gene was cloned into pET28b vector as explained in the previous step. (C) A portion of the gene (two of three repeats) encoding three repeat (Tom20-3) was amplified by PCR using pET28b-containing Tom20-2 gene as a template with Tom20-2-cap_fr and Tom20-1-stop_rv primers. The other portion of the gene (three of three repeats) was produced by PCR with Tom20-3_fr and Tom20-3_rv overlapping primers. Finally, the full gene encoding Tom20-2 gene was produced by an additional PCR reaction with Tom20-3_fr and Tom20-1-stop_rv primers corresponding to the gene termini, and by using the previous two independent PCR reaction products as templates. The gene was cloned into pET28b vector as explained before.

Table 2.1: **Synthetic oligonucleotides used in the construction of Tom20-1, Tom20-2, and Tom20-3 genes.** All the synthetic oligonucleotide sequences are given in standard 5′ to 3′ direction.

Name	**Sequence**
Tom20-1_fr	`GGAATTCCATATGGGCAACAGCTTCCTTGAAGAGATACAGCTTGGTG`
Tom20-1-stop_rv	`CCGCTCGAGTCATTCCACATCATCCTCAGCC`
Tom20-1-cap_fr	`CGCGATTGCGGTGGATCCGAACAACTTCCTTGAAGAGATACAGCTTGGTG`
Tom20-2_fr	`GGAATTCCATATGGGCAACAGCTTTCTGGAAGAAATTCAGCTGGGCGAAGAACTGCTGGCGCAGGGCGATTATG`
Tom20-2_rv	`GTTGTTCGGATCCACCGCAATCGCGTTGGTCAGATGATCCACGCCTTTTTCATAATCGCCCTGCGCCAGCAG`
Tom20-2-cap_fr	`CGAATGCCATCGCCGTTGACCCAAATAATTTTCTGGAAGAAATTCAGCTGGG`
Tom20-3_fr	`GGAATTCCATATGGGTAATTCGTTCTTGGAGGAGATCCAATTGGGTGAGGAGTTGTTGGCCCAAGGTGACTACG`
Tom20-3_rv	`ATTATTTGGGTCAACGGCGATGGCATTCGTCAAGTGGTCAACACCCTTCTCGTAGTCACCTTGGGCCAAC`

2.2.1.2 RPS20-1, RPS20-2 and RPS20-3 genes

Expression plasmids encoding one, two, and three repeat genes were constructed as follows using the primers mentioned in Table 2.2. (A) The gene encoding the single repeat (RPS20-1) was amplified by PCR using genomic DNA from *Thermus thermophilus* as a template with RPS20-1_fr and RPS20-1-stop_rv primers (Table 2.2). The PCR product was purified by agarose gel electrophoresis followed by Quiaquick column purification from agarose gel slices (Gel extraction kit, Quiagen). This product was double digested with NdeI/HindIII and following purification with Quiaquick purification columns (Quiagen) the DNA was ligated to NdeI/HindIII-digested pET28b vector (Invitrogen). Thereby RPS20-1 encoding gene was cloned in frame with the 6×-His tags of the vector plasmid. (B) A portion of the gene (one of two repeats) encoding two repeat (RPS20-2) was amplified by PCR using pET28b-containing RPS20-1 gene as a template with RPS20-1-cap_fr and RPS20-1-stop_rv primers. The other portion of the gene (two of two repeats) was produced by PCR with RPS20-2_fr and RPS20-2_rv overlapping primers. Finally, the full gene encoding RPS20-2 gene was produced by an additional PCR reaction with RPS20-2_fr and RPS20-1-stop_rv primers corresponding to the gene termini, and by using the previous two independent PCR reaction products as templates. The gene was cloned into pET28b vector as explained in the previous step. (C) A portion of the gene (two of three repeats) encoding three repeat (RPS20-3) was amplified by PCR using pET28b-containing RPS20-2 gene as a template with RPS20-2-cap_fr and RPS20-1-stop_rv primers. The other portion of the gene (three of three repeats) was produced by PCR with RPS20-3_fr and RPS20-3_rv overlapping primers. Finally, the full gene encoding RPS20-2 gene was produced by an additional PCR reaction with RPS20-3_fr and RPS20-1-stop_rv primers corresponding to the gene termini, and by using the previous two independent PCR reaction products as templates. The gene was cloned into pET28b vector as explained earlier.

Table 2.2: **Synthetic oligonucleotides used in the construction of RPS20-1, RPS20-2, and RPS20-3 genes.** All the synthetic oligonucleotide sequences are given in standard 5′ to 3′ direction.

Name	**Sequence**
RPS20-1_fr	`GGAATTCCATATGGGCAACAGCATCAAGACCCTCAGCAAGAAGGCC`
RPS20-stop_rv	`CCCAAGCTTTTAGGCGCTGAGGCCGCCGCC`

Continued on next page

Table 2.2 – continued from previous page

Name	Sequence
RPS20-1-cap_fr	GGCCGAGTCGTTGGACCCAAATAACATCAAGACCCTCAGCAAGAAGGCC
RPS20-2_fr	GGAATTCCATATGGGTAATTCGATCAAGACGTTGTCGAAGAAGGCCATCCAATTGGCCCAAGAGGGTAAGGCCG
RPS20-2_rv	CTTATTTGGGTCCAACGACTCGGCCTTACGCATGATCTTCAAGGCCTCCTCGGCCTTACCCTCTTGGGCCAATTGG
RPS20-2-cap_fr	GGCAGAATCCTTAGACCCTAACAACATCAAGACGTTGTCGAAGAAGGCC
RPS20-3_fr	GGAATTCCATATGGGGAACTCCATTAAGACTTTATCCAAGAAGGCAATTCAGTTAGCACAGGAAGGGAAGGCAG
RPS20-3_rv	GTTGTTAGGGTCTAAGGATTCTGCCTTCCGCATAATCTTTAATGCTTCTTCTGCCTTCCCTTCCTGTGCTAACTG

2.2.1.3 PLC-1, PLC-2 and PLC-3 genes

Expression plasmids encoding one, two, and three repeat genes were constructed as follows using the primers mentioned in Table 2.3. (A) The gene encoding the single repeat (PLC-1) was produced by PCR with PLC-1_fr, PLC-1_rv, PLC-1-stop_fr and PLC-1-stop_rv overlapping primers, followed by an additional PCR reaction with PLC-1_fr and PLC-1-stop_rv primers corresponding to the gene termini by using a previous PCR reaction product as a template. The gene was cloned into pET28b vector as explained previously with NdeI/HindIII restriction sites. (B) A portion of the gene (one of two repeats) encoding two repeat (PLC-2) was amplified by PCR using pET28b-containing PLC-1 gene as a template with PLC-1-cap_fr and PLC-1-stop_rv primers. The other portion of the gene (two of two repeats) was produced by PCR using PLC-2_fr and PLC-2_rv overlapping primers. Finally, the full gene encoding PLC-2 gene was produced by an additional PCR reaction with PLC-2_fr and PLC-1-stop_rv primers corresponding to the gene termini, and by using the previous two independent PCR reaction products as templates. The gene was cloned into pET28b vector as explained in the previous step. (C) A portion of the gene (two of three repeats) encoding three repeat (PLC-3) was amplified by PCR using pET28b-containing PLC-2 gene as a template with PLC-2-cap_fr and PLC-1-stop_rv primers. The other portion of the gene (three of three repeats) was produced by PCR with PLC-3_fr and PLC-3_rv overlapping primers. Finally, the full gene encoding PLC-3 gene was produced by an additional PCR reaction with PLC-3_fr and PLC-1-stop_rv primers corresponding to the gene termini, and by using the previous two independent PCR reaction products as templates. The gene was cloned into pET28b vector as explained before.

Table 2.3: **Synthetic oligonucleotides used in the construction of PLC-1, PLC-2, and PLC-3 genes.** All the synthetic oligonucleotide sequences are given in standard 5′ to 3′direction.

Name	**Sequence**
PLC-1_fr	`GGAATTCCATATGGGCAACAGCGTGCGCAAATTTACCGCGCTGGCGCGCAACGAATGGGAAAAAGGC`
PLC-1_rv	`GTTGTTCGGATCATAATGCATCGCCTGGCCAAAATACCAGGTCGCTTTTTCATAGTTGCCTTTTTCCCATTCGTTGCG`
PLC-1-stop_fr	`GATGCATTATGATCCGAACAACGCGGAAGCGAAACAGAACAACCTGGGCAACGC`
PLC-1-stop_rv	`CCCAAGCTTTTAGCCCTGTTTCTGTTTCGCGTTGCCCAGGTTG`
PLC-1-cap_fr	`GCCATGCACTACGACCCAAATAATGTGCGCAAATTTACCGCGC`
PLC-2_fr	`GGAATTCCATATGGGTAATTCGGTTCGTAAGTTCACGGCCTTGGCCCGTAATGAGTGGGAGAAGGGTAATTACG`
PLC-2_rv	`ATTATTTGGGTCGTAGTGCATGGCTTGACCGAAGTACCACGTGGCCTTCTCGTAATTACCCTTCTCCCAC`
PLC-2-cap_fr	`GGGCAGGCAATGCACTACGACCCTAACAACGTTCGTAAGTTCACGGCCTTGGCCCG`
PLC-3_fr	`GAATTCCATATGGGGAACTCCGTCCGGAAGTTCACTGCATTAGCACGGAACGAATGGGAAAAGGGGAACTACG`
PLC-3_rv	`GTTGTTAGGGTCGTAGTGCATTGCCTGCCCGAAGTACCAAGTTGCCTTTTCGTAGTTCCCCTTTTCCCATTCG`

2.2.1.4 HSC-1, HSC-2 and HSC-3 genes

Expression plasmids encoding one, two, and three repeat genes were constructed as follows using the primers mentioned in Table 2.4. (A) The gene encoding the single repeat (HSC-1) was produced by PCR with HSC-1_fr, HSC-1_rv, HSC-1-stop_fr and HSC-1-stop_rv overlapping primers, followed by an additional PCR reaction with HSC-1_fr and HSC-1-stop_rv primers corresponding to the gene termini by using a previous PCR reaction product as a template. The gene was cloned into pET28b vector as explained previously with NdeI/HindIII restriction sites. (B) A portion of the gene (one of two repeats) encoding two repeat (HSC-2) was amplified by PCR using pET28b-containing HSC-1 gene as a template with HSC-1-cap_fr and HSC-1-stop_rv primers. The other portion of the gene (two of two repeats) was produced by PCR using HSC-2_fr and HSC-2_rv overlapping primers. Finally, the full gene encoding HSC-2 gene was produced by an additional PCR reaction with HSC-2_fr and HSC-1-stop_rv primers corresponding to the gene termini, and by using the previous two independent PCR reaction products as templates. The gene was cloned into pET28b vector as explained in the previous step. (C) A portion of the gene (two of three repeats) encoding three repeat (HSC-3) was amplified by PCR using pET28b-containing HSC-2 gene as a template with HSC-2-cap_fr and HSC-1-stop_rv primers. The other portion of the gene (three of three repeats) was produced by PCR with HSC-3_fr and HSC-3_rv overlapping primers. Finally, the full gene encoding HSC-3 gene was produced by an additional PCR reaction with HSC-3_fr and HSC-1-stop_rv primers corresponding to the gene termini, and by using the previous two independent PCR reaction products as templates. The gene was cloned into pET28b vector as explained before.

Table 2.4: **Synthetic oligonucleotides used in the construction of HSC-1, HSC-2, and HSC-3 genes.** All the synthetic oligonucleotide sequences are given in standard 5′ to 3′ direction.

Name	**Sequence**
HSC-1_fr	`GGAATTCCATATGGGCAACAGCATTAAAACCCGCAGCCAGCTGATGGTGCAGCAGCTGGATGAACAGCAGTGG`
HSC-1_rv	`GTTGTTCGGGTCCAGAAAGCGCAGTTTGCGCACGGTATCCGCCGCCTGTTCCCACTGCTGTTCATCCAGCTGC`
HSC-1-stop_fr	`CTGCGCTTTCTGGACCCGAACAACGCGGAAGCGAAACAGAACAACC`
HSC-1-stop_rv	`CCGCTCGAGTCAGCCCTGTTTCTGTTTCGCGTTGCCCAGGTTGTTCTGTTTCGC`
HSC-1-cap_fr	`GCGTTTCTTGGACCCAAATAATATTAAAACCCGCAGCCAGCTGATGGTGC`
HSC-2_fr	`GGAATTCCATATGGGTAATTCGATCAAGACGCGTTCGCAATTGATGGTTCAACAATTGGACGAGCAACAATGGG`

Continued on next page

Table 2.4 – continued from previous page

Name	Sequence
HSC-2_rv	GGTTTTAATATTATTTGGGTCCAAGAAACGCAACTTACGAACCGTGTCGGCGGCTTGCTCCCATTGTTGCTCGTCC
HSC-2-cap_fr	CGGTTCTTAGACCCTAACAACATCAAGACGCGTTCGCAATTGATGG
HSC-3_fr	GGAATTCCATATGGGGAACTCCATTAAGACTCGGTCCCAGTTAATGGTCCAGCAGTTAGACGAACAGCAGTGGGAAC
HSC-3_rv	CGTCTTGATGTTGTTAGGGTCTAAGAACCGTAACTTCCGGACAGTGTCTGCTGCCTGTTCCCACTGCTGTTCGTC

2.2.1.5 BGA-1, BGA-2 and BGA-3 genes

Expression plasmids encoding one, two, and three repeat genes were constructed as follows using the primers mentioned in Table 2.4. (A) The gene encoding the single repeat (BGA-1) was produced by PCR with BGA-1_fr, BGA-1_rv, BGA-1-stop_fr and BGA-1-stop_rv overlapping primers, followed by an additional PCR reaction with BGA-1_fr and BGA-1-stop_rv primers corresponding to the gene termini by using a previous PCR reaction product as a template. The gene was cloned into pET28b vector as explained previously with NdeI/HindIII restriction sites. (B) A portion of the gene (one of two repeats) encoding two repeat (BGA-2) was amplified by PCR using pET28b-containing BGA-1 gene as a template with BGA-1-cap_fr and BGA-1-stop_rv primers. The other portion of the gene (two of two repeats) was produced by PCR using BGA-2_fr and BGA-2_rv overlapping primers. Finally, the full gene encoding BGA-2 gene was produced by an additional PCR reaction with BGA-2_fr and BGA-1-stop_rv primers corresponding to the gene termini, and by using the previous two independent PCR reaction products as templates. The gene was cloned into pET28b vector as explained in the previous step. (C) A portion of the gene (two of three repeats) encoding three repeat (BGA-3) was amplified by PCR using pET28b-containing BGA-2 gene as a template with BGA-2-cap_fr and BGA-1-stop_rv primers. The other portion of the gene (three of three repeats) was produced by PCR with BGA-3_fr and BGA-3_rv overlapping primers. Finally, the full gene encoding BGA-3 gene was produced by an additional PCR reaction with BGA-3_fr and BGA-1-stop_rv primers corresponding to the gene termini, and by using the previous two independent PCR reaction products as templates. The gene was cloned into pET28b vector as explained before.

Table 2.5: **Synthetic oligonucleotides used in the construction of BGA-1, BGA-2, and BGA-3 genes.** All the synthetic oligonucleotide sequences are given in standard 5′ to 3′ direction.

Name	Sequence
BGA-1_fr	`GGAATTCCATATGGGCAACAGCAGCCGCGATCTGTATCATGTGGCGAACGCGTTTATTGCGGCGGGCGATGTG`
BGA-1_rv	`GCTTCCGCGTTGTTCGGGTCCACTTTCGCCAGATAATCCAGGCTGCGGTTCGCGCTATCCACATCGCCCGCCGC`
BGA-1-stop_fr	`GACCCGAACAACGCGGAAGCGAAACAGAACAACCTGGGCAACGCG`
BGA-1-stop_rv	`CCGCTCGAGTCAGCCCTGTTTCTGTTTCGCGTTGCCCAGGTTG`
BGA-1-cap_fr	`GCCAAGTTGACCCAAATAATAGCCGCGATCTGTATCATG`
BGA-2_fr	`GGAATTCCATATGGGTAATTCGTCGCGTGACTTGTACCACGTTGCCAATGCCTTCATCGCCGCCGG`
BGA-2_rv	`GCTATTATTTGGGTCAACCTTGGCCAAGTAGTCCAACGAACGATTGGCCGAGTCAACGTCACCGGCGGCGATGAAG`
BGA-2-cap_fr	`GCAAAGGTCGACCCTAACAACTCGCGTGACTTGTACCACGTTGC`
BGA-3_fr	`GGAATTCCATATGGGGAACTCCTCCCGGGACTTATACCACGTCGCAAACGCATTCATTGCAGCAGGGG`
BGA-3_rv	`CGAGTTGTTAGGGTCGACCTTTGCTAAGTAGTCTAAGGACCGGTTTGCGGAGTCGACGTCCCCTGCTGCAATGAATGCG`

2.2.2 Transformation

E. coli TOP10 cells were grown in Luria Bertani (LB) medium to the early exponential phase ($A_{600}\sim$ 0.35- 0.37), and then pelleted by centrifugation at 1000 × g for 10 min at 4 °C. After that resuspended at one-tenth of their original volume in ice-cold sterial $CaCl_2$ solution, and incubated for 2 hours at 4 °C. A 50 μl aliquot of cells were transferred into a cold polypropylene tube, mixed with each of the ligation reaction mixture, and incubated for 30 minute on ice. Next, the cells were heat-shocked for exactly 45 seconds in the 42 °C water bath, and were quickly placed on ice for 2 minutes. The 250 μl of pre-warmed (37 °C) LB medium was added to each vial, and the cells were grown at 37 °C with shaking

(225 rpm) for 1 hour to allow expression of the antibiotic-resistance gene. Transformants were selected by plating cells on LB agar plates containing 50 μg/ml kanamycin, after incubation of the plates at 37 °C for 18-20 hours.

2.2.3 Plasmid DNA extraction

Liquid cultures of kanamycin-resistant bacteria were prepared by innoculating 4 ml aliquots of LB medium with either single bacterial colonies or frozen bacterial glycerol stocks. Plasmid DNA was extracted from saturated overnight cultures using a column-based extraction system according to the manufacturer's instructions (Mini prep, Quiagen) and the final DNA pellets were dissolved in 50μl of Millipore filtered, glass distilled water. Yields of DNA was calculated based on optical density measurements of dilutions made from the stocks. The optical density was measured at both 260nm and 280nm which is later used to assess DNA purity as well as concentration. Glycerol stocks were prepared for each plasmid produced according to standard methods.

2.2.4 DNA sequencing

Sequencing reactions were performed on an Applied Biosystem DNA Sequencer by the dideoxy chain termination method. Taq sequencing was performed at the central sequencing service of the Max-Planck Institute for developmental biology, using T7 promotor or T7 terminator primer. The following temperature cycling parameters were used to perform the reactions with the ABI Prism "Big Dye" Terminator Cycle sequencing kit: 96 °C/20 seconds, 50 °C/10 seconds and 60 °C/4 minutes for 30 cycles.

2.2.5 Protein production

All the fifteen constructs in pET28b vector were transformed into *Escherichia coli* expression host BL21-Gold (DE3) (Stratagene) strain. Constructs were tested for expression in a small volume (5 ml) prior to expression on large scale in LB medium. Cultures were grown in LB media at 37 °C with constant agitation (270 rpm). After approximately 2 hr (A_{600}~0.6), expression was induced by adding isopropyl β-D-thiogalactopyranoside (IPTG) to a final concentration of 1 mM. Cultures were grown for an additional 4 hr, and the cells were harvested by centrifugation (15 min, 2,500 × g, 5 °C). Finally, the pellets were stored at -80 °C for later purification.

2.2.6 Protein concentration determination

Protein concentration was determined using three different methods. For most of the pure protein samples, ultraviolet absorption of the proteins at 280 nm was used. Extinction coefficients for proteins was calculated using ProtParam software (`http://www.expasy.ch/tools/protparam.html`) and protein concentration was calculated from the measured absorbance at 280 nm.

Using 1 mg/ml trypsin stock solution 0, 0.2, 0.4, 0.6, 0.8, and 1 mg/ml concentrations were taken in separate cuvettes. Similarly each sample was taken in triplicate. The volume of each cuvette was made upto 50 μl using corresponding buffer. 1 ml of a freshly prepared BCA solution (solution A : solution B = 50:1) was added to each cuvette and then incubated for 30 min at 37 °C. Subsequently, the protein content was measured by the absorbance at 562 nm, where the protein concentration was determined from the standard curve equation.

To follow the purification of the proteins, the modified Bradford assay called drop assay was used. The 10 μl of fractions were mixed with 30 μl of the Bradford reagent in a drop on a clean stripe of parafilm. Steady blue color was used as a sign for the protein presence in the tested fractions, which were then further analyzed by tricine-urea-SDS-PAGE.

2.2.7 Tricine-urea-SDS-PAGE

Tricine-urea-SDS-PAGE was performed using 15 % gels (8 x 10 x 0.75 cm) according to Von Jagow *et al.* [136] method. Gels were prepared by following the standard protocol. Samples were mixed with 4× SDS sample buffer and boiled at 95 °C for 5 min before loading onto the gel. The low molecular weight marker kit from GE-Healthcare was used to estimate the protein molecular weight. The protein names and molecular weights are tabulated in Table 2.6. Electrophoresis was performed using electrophoresis buffer at a constant current of 23 mA per gel.

Table 2.6: **Molecular weight marker proteins.** The list of standard molecular weight marker proteins and their sizes used in the tricine-urea-SDS-PAGE.

Protein	**Mol. Wt. (kDa)**
Phosphorylase b	97
Albumin	66
Ovalbumin	45
Carbonic anhydrase	30
Trypsin inhibitor	20.1
α-lactalbumin	14.4

The gels were then incubated in fixative solution for 5-10 min on a shaker, stained with Coomassie staining solution for 15 min and destained until the protein bands became clearly visible. The gels are photographed for documentation.

2.2.8 Protein purification

Frozen cell pellets were thawed and resuspended in 20-30 ml of ice-cold lysis buffer (30 mM $NaH_2PO_4.2H_2O$, 50 mM NaCl, 2 mM $MgCl_2$, a pinch of DNAse I, pH 7.2) and ruptured using a French press. The soluble fraction or inclusion bodies were separated by centrifugation (30 min, 15,000 × g, 5 °C). The soluble proteins were purified under native conditions, whereas insoluble ones were purified under denaturing conditions.

2.2.8.1 Tom20-1, Tom20-2 and Tom20-3 proteins

The following steps for purification were identical for Tom20-1, Tom20-2 and Tom20-3 proteins. According to manufacturer's instructions, the soluble fraction of the lysate was loaded onto a column of Ni Sepharose 6 Fast Flow (10ml, GE Healthcare). The fractions containing Tom20-1/2/3 proteins were combined, and loaded onto a second Superdex G-75 26/60 gel-size exclusion chromatography column (GE Healthcare) and eluted with Phosphate-buffered saline (PBS).

2.2.8.2 RPS20-1, RPS20-2 and RPS20-3 proteins

The following steps for purification were identical for RPS20-1, RPS20-2 and RPS20-3 proteins. The supernatant (20-30 ml) was loaded onto a column (10 by 26 cm) of Hi Load 26/10 SP Sepharose High Performance (GE Healthcare) equilibrated with wash buffer (30 mM $NaH_2PO_4.2H_2O$, 50 mM NaCl, pH 8.5). The column was eluted with a linear gradient of 0 to 1.0 M NaCl in the elution buffer (30 mM $NaH_2PO_4.2H_2O$, 1.0 M NaCl, pH 8.5). According to manufacturer's instructions, the fractions containing RPS20-1/2/3 proteins were combined, and loaded onto a second column of Ni Sepharose 6 Fast Flow (10ml, GE Healthcare). Finally, proteins were purified using Superdex G-75 26/60 gel-size exclusion chromatography column (GE Healthcare) by eluting with elution buffer(30 mM $NaH_2PO_4.2H_2O$, 50 mM NaCl, pH 7.2).

2.2.8.3 BGA-1, BGA-2 and BGA-3 proteins

The following steps for purification were identical for BGA-1, BGA-2 and BGA-3 proteins. According to manufacturer's instructions, the soluble fraction of the lysate was loaded onto a column of Ni Sepharose 6 Fast Flow (10ml, GE Healthcare). The fractions containing Tom20-1/2/3 proteins were combined, and loaded onto a second Superdex G-75 26/60 gel-size exclusion chromatography column (GE Healthcare) and eluted with Phosphate-buffered saline (PBS).

2.2.8.4 Purification under denaturing conditions and refolding

The following steps for purification and refolding were identical for PLC-1, PLC-2, PLC-3, HSC-1, HSC-2 and HSC-3 proteins. Inclusion bodies recovered from cell lysate by centrifugation were washed two times with wash buffer (0.1 M Tris/HCl, 2 M urea, 5 mM EDTA, 5 % triton X-100, pH 7.0). Clarified inclusion bodies were solubilized in 8 M urea, and insoluble material was removed by centrifugation. Solubilized inclusion bodies were purified using His Tag-nickel affinity chromatography. The pure proteins were refolded by dialyzing against refolding buffer (30 mM $NaH_2PO_4.2H_2O$, 50 mM NaCl, pH 5.5).

2.2.9 Circular Dichorism (CD)

CD data were collected on a Jasco J-810 spectropolarimeter. Five far-UV scans were averaged, and the buffer base line was subtracted from all spectra registered. Experiments were carried out at 22 °C in a 1 mm pathlength cuvette, at various protein concentrations. Where indicated, 2,2,2-trifluoroethanol (TFE) was added. Temperature-induced protein denaturation was followed by the change in ellipticity at 222 nm in a 10 mm pathlength cuvette. Temperature was varied using a peltier device.

2.2.10 Fluorescence spectroscopy

Fluorescence spectra of tryptophan (Trp) were measured with a Jasco FP-6500 spectrometer, with both emission and excitation band width set at 3 nm. The excitation wavelength was set at 293 nm, and the tryptophan emission was monitored between 300-400 nm. Five scans were averaged and the buffer base line was substracted from all spectra registered. The experiments were carried out at room temperature and in quartz cuvettes with 1 cm pathlength. Where indicated, TFE was added.

2.2.11 Chaperone Assays

Chaperone activity of Tom20-1/2/3 proteins was tested in heat-induced protein aggregation assays, with porcine citrate synthase (Sigma) as a substrate. Citrate synthase (0.3 μM) was mixed with the respective Tom20-1/2/3 protein in fivefold excess molar ratios in assay buffer (30 mM Hepes, pH 7.3, 5 mM MgCl2, and 100 mM KCl). Thermal aggregation was monitored at 45 °C as increase in attenuance at 340 nm in a Perkin-Elmer Lambda 25 UV/VIS Spectrometer equipped with a thermostatted cuvette holder. All results were normalized relative to the heat-induced denatured citrate synthase subject to the same conditions.

Chapter 3

Results

3.1 Bioinformatics analysis

3.1.1 Structure-based sequence alignment

The structure of each TPR unit is virtually identical; first repeat of protein phosphatase 5 (1A17, chain A) superimposes with all other 33 individual repeats with a 2.6 $\AA$ maximum root mean square deviation (RMSD) as computed from the Swiss-PDB Viewer. Figure 3.1A shows the structure-based sequence alignment, and reveals a high degree of sequence diversity within the repeats. Although there is no position characterised by an invariant residue, a consensus sequence pattern of aromatic and hydrophobic residues has been observed. Hydrophobic residues are commonly observed at positions 1, 12, 20, 21, 27 and 28, while glycine and proline frequently occupies positions 8 and 32 respectively within the repeat. Aromatic residues like tyrosine, tryptophan or phenylalanine are frequent at position 17.

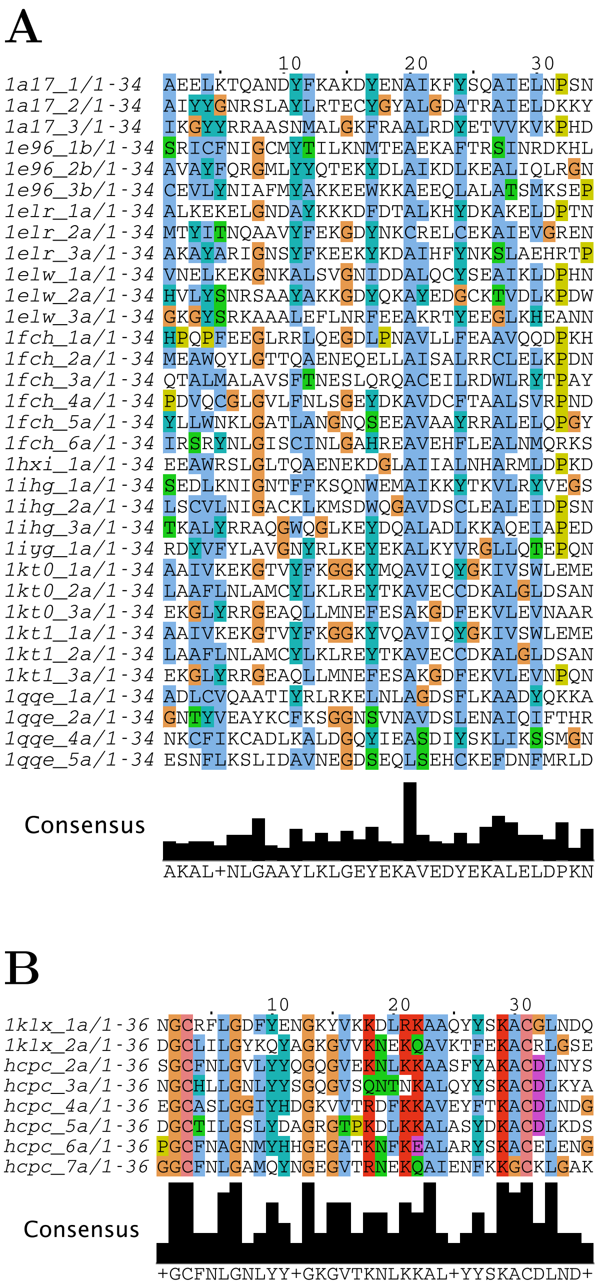
The repeating units of the TPR superimpose closely (less than 3 $\AA$ RMSD) with equivalent repeats of SEL1-like repeats with the insertion of four residues in the loop region that connects the two helices of the repeat as illustrated in Figure 3.1. Figure 3.1B shows the structure-based sequence alignment of SEL1-like repeats, and reveals that 11 out of 36 residues are conserved in almost all repeats. Glycine residues at positions 2, 7 and 13, cysteine residues at positions 3 and 31, leucine residues at positions 6 and 33, alanine residues at positions 23 and 30 and lysine residues at positions 18 and 29 are generally conserved. Unlike TPR repeating units, SEL1-like repeats are held together by disulfide bridges between highly conserved cysteine pairs at position 3 and 31. TPR-containing proteins and SEL1-like repeat proteins share common structural and functional properties, despite their lack of sequence similarities.

The philosophy behind TPRpred is to construct a seed alignment for each family of repeats from a representative non-redundant set of domain sequences trusted to belong to the family. Therefore, the structural alignments shown in Figure 3.1 were used as a seed alignment for TPR and SEL1-like repeat families. From each of the seed alignments the profile was built, which was subsequently used to generate series of profiles.

3.1.2 Profile generation and selection

The selectivity of sequence profiles depends on the number of close homologs, whereas the sensitivity depends on the number of remote homologs used in constructing the profiles. Relaxing the threshold value to include remote homologs often results in false positives. To optimize the trade-off between remote homologs and false positives, we have constructed a series of TPR profiles. These profiles were generated by iterative searches against the non-redundant (NR) database at NCBI (www.ncbi.nlm.nih.gov), filtered to a 70 % maximum sequence identity (NR-70) by CD-HIT [138, 139]. Prior to the searches we broadly removed homologs of MalT [GI:16131294], which we intended to use as a test set, from the NR-70 database using three iterations of PSI-BLAST [52] at an E-value cutoff of 1.

Homologs of MalT contain divergent TPR units and therefore represent a challenging test set. These proteins belong to the STAND family of ATPases [140, 141], which themselves are part of the AAA+ superfamily [142]. We extracted these sequences conservatively with PSI-BLAST (two iterations, E-value cutoff of 10^{-4}) from NR-70, using the central domain of MalT [GI:17942835] as a query sequence. Using the defining characteristic of STAND proteins, namely an N-terminal P-loop NTPase domain, as a criterion we selected 56 proteins for the test set.



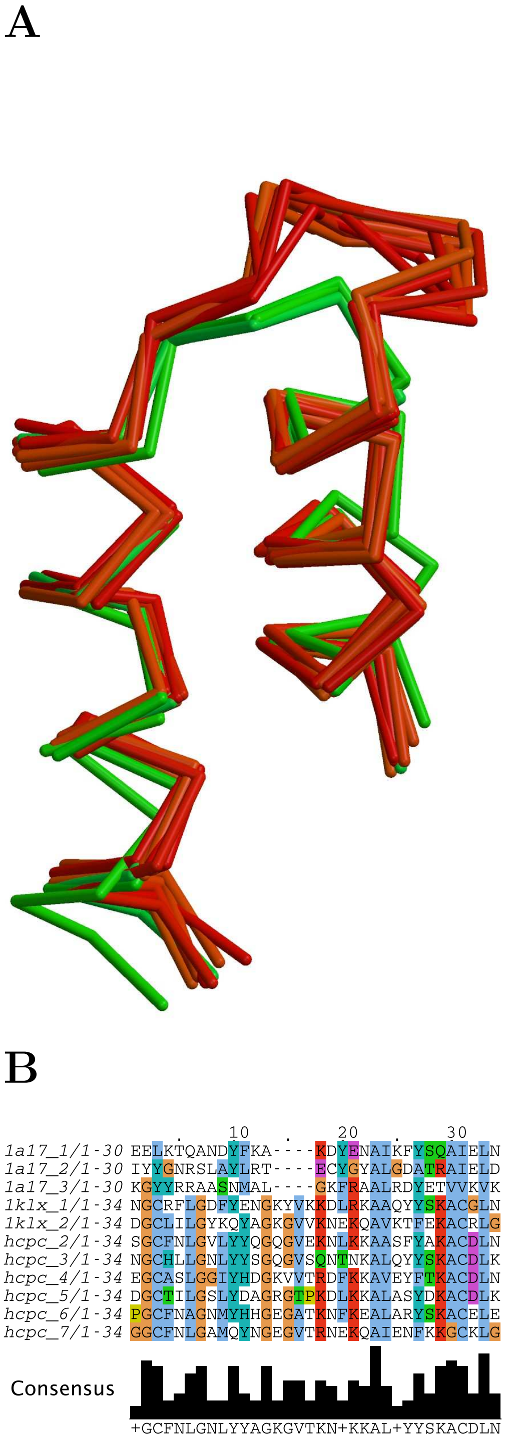
```
                      10    .    20    .    30
1a17_1/1-34    AEELKTQANDYFKAKDYENAIKFYSQAIELNPSN
1a17_2/1-34    AIYYGNRSLAYLRTECYGYALGDATRAIELDKKY
1a17_3/1-34    IKGYYRRAASNMALGKFRAALRDYETVVKVKPHD
1e96_1b/1-34   SRICFNIGCMYTILKNMTEAEKAFTRSINRDKHL
1e96_2b/1-34   AVAYFQRGMLYYQTEKYDLAIKDLKEALIQLRGN
1e96_3b/1-34   CEVLYNIAFMYAKKEEWKKAEEQLALATSMKSEP
1elr_1a/1-34   ALKEKELGNDAYKKKDFDTALKHYDKAKELDPTN
1elr_2a/1-34   MTYITNQAAVYFEKGDYNKCRELCEKAIEVGREN
1elr_3a/1-34   AKAYARIGNSYFKEEKYKDAIHFYNKSLAEHRTP
1elw_1a/1-34   VNELKEKGNKALSVGNIDDALQCYSEAIKLDPHN
1elw_2a/1-34   HVLYSNRSAAYAKKGDYQKAYEDGCKTVDLKPDW
1elw_3a/1-34   GKGYSRKAAALEFLNRFEEAKRTYEEGLKHEANN
1fch_1a/1-34   HPQPFEEGLRRLQEGDLPNAVLLFEAAVQQDPKH
1fch_2a/1-34   MEAWQYLGTTQAENEQELLAISALRRCLELKPDN
1fch_3a/1-34   QTALMALAVSFTNESLQRQACEILRDWLRYTPAY
1fch_4a/1-34   PDVQCGLGVLFNLSGEYDKAVDCFTAALSVRPND
1fch_5a/1-34   YLLWNKLGATLANGNQSEEAVAAYRRALELQPGY
1fch_6a/1-34   IRSRYNLGISCINLGAHREAVEHFLEALNMQRKS
1hxi_1a/1-34   EEAWRSLGLTQAENEKDGLAIIALNHARMLDPKD
1ihg_1a/1-34   SEDLKNIGNTFFKSQNWEMAIKKYTKVLRYVEGS
1ihg_2a/1-34   LSCVLNIGACKLKMSDWQGAVDSCLEALEIDPSN
1ihg_3a/1-34   TKALYRRAQGWQGLKEYDQALADLKKAQEIAPED
1iyg_1a/1-34   RDYVFYLAVGNYRLKEYEKALKYVRGLLQTEPQN
1kt0_1a/1-34   AAIVKEKGTVYFKGGKYMQAVIQYGKIVSWLEME
1kt0_2a/1-34   LAAFLNLAMCYLKLREYTKAVECCDKALGLDSAN
1kt0_3a/1-34   EKGLYRRGEAQLLMNEFESAKGDFEKVLEVNAAR
1kt1_1a/1-34   AAIVKEKGTVYFKGGKYVQAVIQYGKIVSWLEME
1kt1_2a/1-34   LAAFLNLAMCYLKLREYTKAVECCDKALGLDSAN
1kt1_3a/1-34   EKGLYRRGEAQLLMNEFESAKGDFEKVLEVNPQN
1qqe_1a/1-34   ADLCVQAATIYRLRKELNLAGDSFLKAADYQKKA
1qqe_2a/1-34   GNTYVEAYKCFKSGGNSVNAVDSLENAIQIFTHR
1qqe_4a/1-34   NKCFIKCADLKALDGQYIEASDIYSKLIKSSMGN
1qqe_5a/1-34   ESNFLKSLIDAVNEGDSEQLSEHCKEFDNFMRLD

Consensus
               AKAL+NLGAAYLKLGEYEKAVEDYEKALELDPKN
```

```
                     10    .    20    .    30
1klx_1a/1-36  NGCRFLGDFYENGKYVKKDLRKAAQYYSKACGLNDQ
1klx_2a/1-36  DGCLILGYKQYAGKGVVKNEKQAVKTFEKACRLGSE
hcpc_2a/1-36  SGCFNLGVLYYQGQGVEKNLKKAASFYAKACDLNYS
hcpc_3a/1-36  NGCHLLGNLYYSGQGVSQNTNKALQYYSKACDLKYA
hcpc_4a/1-36  EGCASLGGIYHDGKVVTRDFKKAVEYFTKACDLNDG
hcpc_5a/1-36  DGCTILGSLYDAGRGTPKDLKKALASYDKACDLKDS
hcpc_6a/1-36  PGCFNAGNMYHHGEGATKNFKEALARYSKACELENG
hcpc_7a/1-36  GGCFNLGAMQYNGEGVTRNEKQAIENFKKGCKLGAK

Consensus
              +GCFNLGNLYY+GKGVTKNLKKAL+YYSKACDLND+
```

Figure 3.1: **Structure-based sequence alignments**. (A) Structural alignment of 33 individual TPR units from the 11 TPR-containing structures. (B) Structural alignment of 8 individual SEL1-like repeat units from 2 SEL1-like repeat-containing structures. The lower panel of each alignment shows the consensus histogram. The figure was prepared using JalView [137].



```
                     10        20        30
1a17_1/1-30 EELKTQANDYFKA----KDYENAIKFYSQAIELN
1a17_2/1-30 IYYGNRSLAYLRT----ECYGYALGDATRAIELD
1a17_3/1-30 KGYYRRAASNMAL----GKFRAALRDYETVVKVK
1klx_1/1-34 NGCRFLGDFYENGKYVKKDLRKAAQYYSKACGLN
1klx_2/1-34 DGCLILGYKQYAGKGVVKNEKQAVKTFEKACRLG
hcpc_2/1-34 SGCFNLGVLYYQGQGVEKNLKKAASFYAKACDLN
hcpc_3/1-34 NGCHLLGNLYYSGQGVSQNTNKALQYYSKACDLK
hcpc_4/1-34 EGCASLGGIYHDGKVVTRDFKKAVEYFTKACDLN
hcpc_5/1-34 DGCTILGSLYDAGRGTPKDLKKALASYDKACDLK
hcpc_6/1-34 PGCFNAGNMYHHGEGATKNFKEALARYSKACELE
hcpc_7/1-34 GGCFNLGAMQYNGEGVTRNEKQAIENFKKGCKLG

            +GCFNLGNLYYAGKGVTKN+KKAL+YYSKACDLN
```

Figure 3.2: **Structure-based sequence alignments of TPRs and SEL1-like repeats.** (A) Superposition of TPRs and SEL1-like repeats. TPRs are shown in green wherease SEL1-like repeats are shown in red color. (B) Structure-based sequence alignment displayed using JalView [137]. The lower panel of an alignment shows the consensus histogram.

We derived two further test sets from proteins of known structure; using the domain sequences of the SCOP database version 1.69 filtered to a 70 % maximum sequence identity, available at ASTRAL [143]. The domain sequences which are classified under TPR family [SCOP:a.118.8.1] and mitochondrial import receptor Tom20 family [SCOP:a.23.4.1] [126] were excluded to generate a true negative set of 9518 domain sequences. In order to make this set more conservative, we submitted the 13 TPR structures from SCOP (a.118.8.1 and a.23.4.1), along with MalT and 5 new TPR domains that were not included in the latest SCOP database version 1.69, to the DALI structure comparison server [66], and excluded all structures with Z scores ≥ 5 from the true negative set (Table 3.1). The 19 TPR-containing structures submitted to the DALI server constitute our the true positive set (TPR structure set).

We performed iterative searches to convergence on NR-70 minus STAND proteins with various threshold parameters (whole-protein E-value, and single-repeat P-value). The initial searches were seeded with a manually prepared structure-based sequence alignment of known TPR protein structures (Figure 3.1A). We tested the resulting profiles on the STAND family, TPR family, and the true negative set. The best profile is selected based on its performance on the STAND family, as illustrated in Figure 3.3.

Further, we built the PPR and the SEL1-like profiles by using the same procedure and cutoff value as for the TPR profile.

Table 3.1: **Structural neighbours of TPRs.** Structural neighbours of known TPRs according to the DALI structure comparison server. The structures with Z scores ≥ 5 are tabulated. The PDB codes were mapped onto the SCOP domain database.

SCOP-ID	Family	Description
d1qoja_	a.2.9.1	C-terminal UvrC-binding domain of UvrB
d1cuna1	a.7.1.1	Spectrin alpha chain
d1owaa_	a.7.1.1	Spectrin alpha chain
d1qlaa1	a.7.3.1	Fumarate reductase

Continued on next page

Table 3.1 – continued from previous page

SCOP-ID	Family	Description
d1hx1b_	a.7.7.1	BAG-family molecular chaperon regulator-1, BAG1
d1m62a_	a.7.7.1	Silencer of death domains, Sodd (Bag4)
d1fjgt_	a.7.6.1	Ribosomal protein S20
d1sumb_	a.7.12.1	PhoU homolog TM1734
d1a32__	a.16.1.2	Ribosomal protein S15
d1mtyg_	a.23.3.1	Methane monooxygenase hydrolase, gamma subunit
d256ba_	a.24.3.1	Cytochrome b562
d1nzea_	a.24.18.1	Oxygen-evolving enhancer protein 3
d2a0b__	a.24.10.1	Aerobic respiration control sensor protein, ArcB
d1ug7a_	a.24.24.1	Domain from hypothetical 2610208m17rik protein
d1ile_1	a.27.1.1	Isoleucyl-tRNA synthetase (IleRS)
d1h3na1	a.27.1.1	Leucyl-tRNA synthetase (LeuRS)
d1rj1a_	a.29.6.1	Invertase inhibitor
d1bg1a1	a.47.1.1	STAT3b
d1uura1	a.47.1.1	STAT homologue coiled coil domain
d1fioa_	a.47.2.1	Sso1
d1hs7a_	a.47.2.1	Vam3p N-terminal domain
d1o5ha_	a.191.1.1	Hypothetical protein TM1560
d1iqpa1	a.80.1.1	Replication factor C
d1h12a_	a.102.1.2	Endo-1,4-beta-xylanase
d1fp3a_	a.102.1.3	N-acyl-D-glucosamine 2-epimerase
d1lf6a1	a.102.1.5	Bacterial glucoamylase, C-terminal domain
d1qaza_	a.102.3.1	Alginate lyase A1-III
d2sqca1	a.102.4.2	Squalene-hopene cyclase
d1qgra_	a.118.1.1	Importin beta
d1qbkb_	a.118.1.1	Karyopherin beta2
d1ee4a_	a.118.1.1	Karyopherin alpha
d1gw5a_	a.118.1.10	Adaptin alpha C subunit N-terminal fragment
d1b3ua_	a.118.1.2	Constant regulatory domain of protein phosphatase 2a
d1uw4b_	a.118.1.14	Regulator of nonsense transcripts 2, UPF2
d1h6ka1	a.118.1.14	CBP80, 80KDa nuclear cap-binding protein
d1hs6a1	a.118.1.7	Leukotriene A4 hydrolase C-terminal domain
d1b89a_	a.118.1.3	Clathrin heavy chain proximal leg segment

Continued on next page

Table 3.1 – continued from previous page

SCOP-ID	Family	Description
d1rz4a2	a.118.1.18	Eukaryotic translation initiation factor 3 subunit 12
d1qsaa1	a.118.5.1	70 KDa soluble lytic transglycosylase (SLT70)
d1klxa_	a.118.18.1	Cysteine rich protein B (HcpB)
d1ouva_	a.118.18.1	Cysteine rich protein C (HcpC)
d1dcea1	a.118.6.1	Rab geranylgeranyltransferase alpha-subunit, N-terminal
d1ld8a_	a.118.6.1	Protein farnesyltransferase alpha-subunit
d1vdua_	a.118.20.1	Hypothetical protein ST1625
d1dtoa_	b.91.1.1	E2 regulatory, transactivation domain
d1ciy_3	f.1.3.1	delta-Endotoxin (insectocide), N-terminal domain
d1i5pa3	f.1.3.1	delta-Endotoxin (insectocide), N-terminal domain

3.1.3 Analysis of the profiles

In a HMM profile, every position consists of a match, insert and delete state, where each position is connected to the next position's states through transitions of altering probability. HMM Logos try to provide a quick overview of the features of a HMM profile while conserving as much information as possible. The height of the stack represents the information content of the distribution of the emission probabilities within some state relative to the background distribution given for the whole profile. The relative size of a letter then expresses it's emission probability from a state's distribution as illustrated in Figure 3.4. The residues are sorted in a descending way depending on their probability. Apart from the white columns, there are pink ones as well. These will most likely not contain any residues: these are the insert states, which normally have an emission distribution very close or equal to the background. Thus, the information content is usually so small that no residues can be seen.

HMM Logo of TPR was constructed from 7679 individual repeat units. This reveals that preference of hydrophobic residue leucine at position 4, 7, 11, 14, 21, 28, 30, alanine at position 8, 20, 27, glycine at position 8, 15, tyrosine at position 11, 24, and proline at position 32 as illustrated in Figure 3.4A. Analysis of the

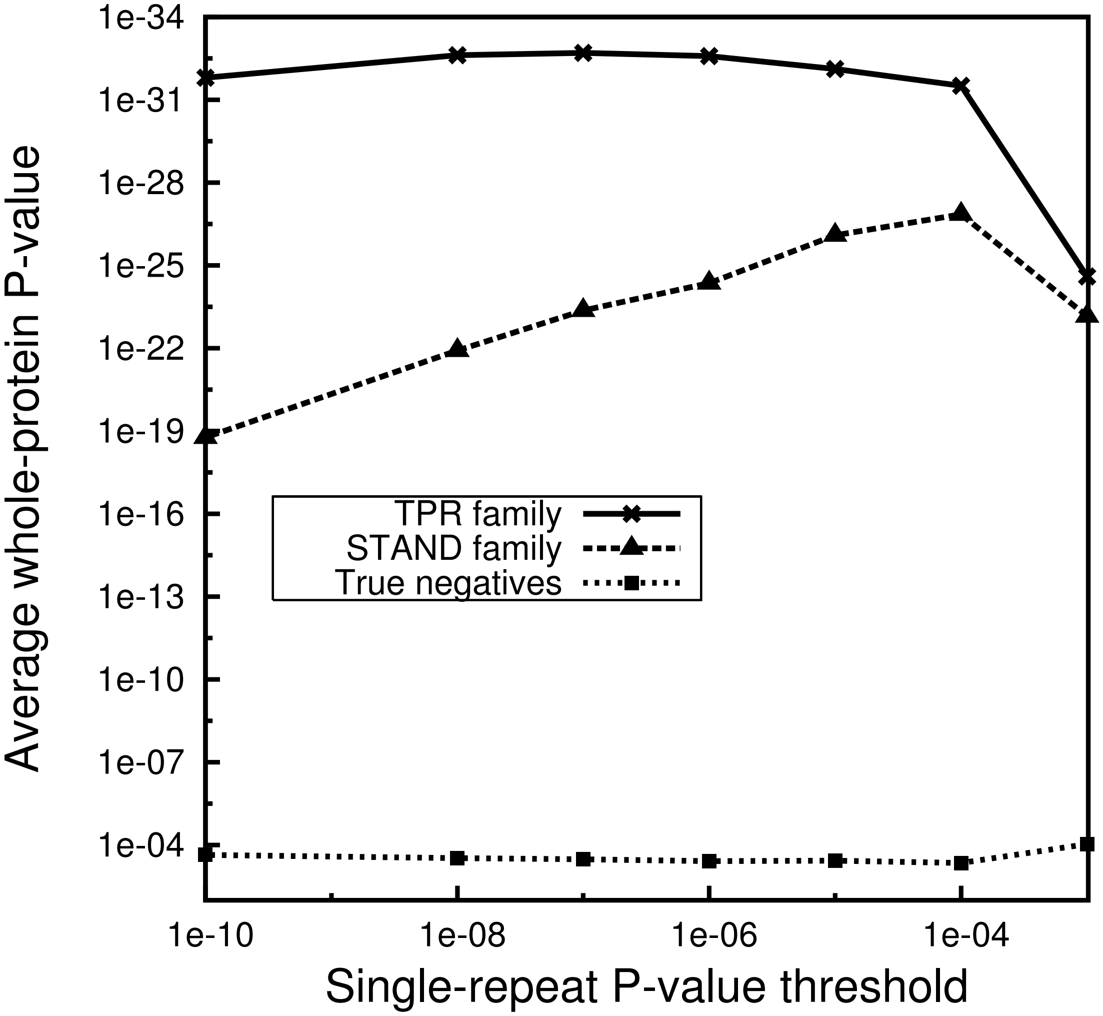


Figure 3.3: **Selection of the best profile**. The geometric average of the whole-protein P-value for the top 10 hits in each test set is plotted against the profile's single-repeat P-value threshold. The profile obtained for a single-repeat P-value threshold of 10^{-4} was selected as best.

structure of the TPR-containing domains provides a rationale for the residue conservation. Residues at position 8 and 20 are located at the position of closest contact between the A and B α-helices of a TPR unit, whereas residue 27 on helix B is located at the interface of 3 helices (A, B and A$'$). Proline 32 is located at the C-terminus of helix B, and the large consensus hydrophobic residues form the interfaces between adjacent α-helices.

The PPRs are named after the characteristic 35-amino-acid motif that constitutes the repeat unit [108]. PPR-containing proteins are essential for RNA processing in orgenelles of higher eukaryotes [110]. There is no 3D structure available for this type of repeats. However, the sequence similarity to the TPR-related repeats and the predicted secondary structure suggest that PPR domain would adopt an α-solenoid structure [144]. HMM Logo of PPR was constructed from 7121 individual repeat units, which reveals that preference of hydrophobic residue leucine at position 6, 7, 10, 16, 22, 26, alanine at position 9, 11, 19, 20, isoleucine at position 6, 7, 22, methionine at position 26, glycine at position 14, 30, tyrosine at position 3, 10, and proline at position 34 as illustrated in Figure 3.4B.

The SEL1-like repeats have 36-amino-acid $\alpha\alpha$-hairpin repeat units. These repeats were first detected in a receptor protein called SEL1 of *C. elegans*, which is a key negative regulator of the Notch pathway. A growing body of evidences suggests that the SEL1-like repeats are involved in protein-protein interactions [109]. HMM Logo of SEL1-like repeats was constructed from 1059 individual repeat units, which reveals that preference of hydrophobic residue leucine at position 5, 8, 9, alanine at position 1, 6, 22, 29, 30, glycine at position 6, 12, 14, 33, tyrosine at position 9, 25, 26, tryptophan at position 25, and aspartic acid at position 18 as shown in Figure 3.4C.

3.1.4 Benchmarking

We benchmarked our method and the web-server against Pfam, SMART and REP.

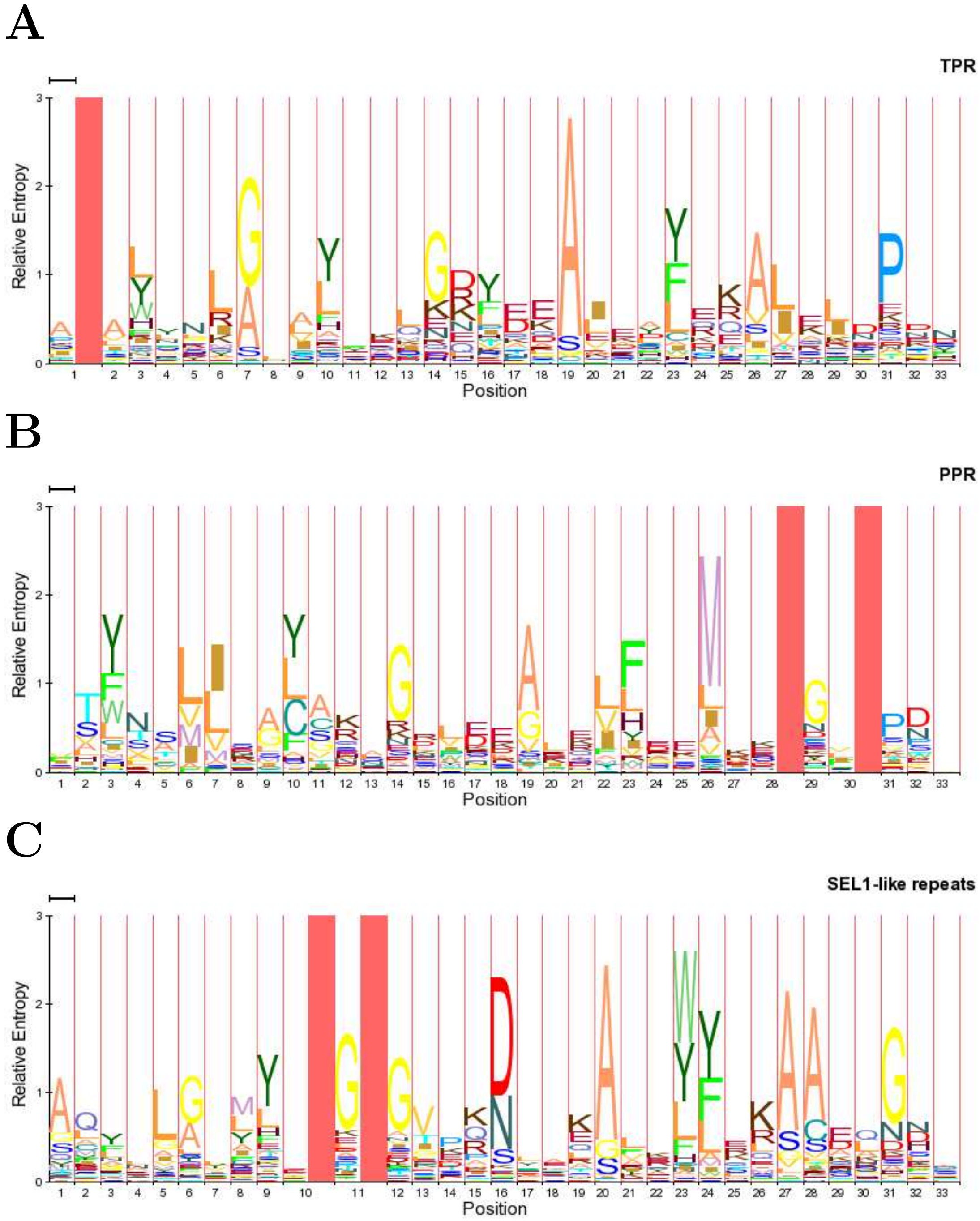


Figure 3.4: **The HMM Logos visualization of the best profiles**. The numbers on X-axis in each logo represents the state positions in the HMM. The overall height of the letter stacks represents the information content, the relative letter height corresponds to its emission probability. Insert states are shown in pink [131].

3.1.4.1 Comparison of TPRpred and HMMER

To demonstrate the sensitivity/selectivity of TPRpred against HMMER (version 2.3), which is the underlying method employed by the Pfam and SMART web-servers, we benchmarked the performance of both these methods, and the results are shown using the receiver operating characteristic (ROC) plot as illustrated in Figure 3.5. We could not benchmark against REP, because the stand-alone version is not available. The data sets for the benchmark were obtained using the same true positive and true negative sets which we defined in the profile generation section, but with a 25 % maximum sequence identity. In order to enrich these data sets with reliable homologs, two iterations of PSI-BLAST searches were performed for each domain sequence. The first iteration was performed on the NR-90 database. The hits with an E-value $\leq 10^{-3}$ and ≥ 85 % coverage to the query sequence were extracted into a multiple alignment, that was used to jump-start the second iteration against the NR-70 database. The same selection criteria as in the first iteration were applied in obtaining the homologs for the query. The resulting enriched data sets were simultaneously filtered to a 50 % maximum sequence identity using CD-HIT to reduce the redundancy.

Both methods were used to perform searches through the true positive and true negative data sets, using their own TPR profiles or HMMs. The ROC plot shows that TPRpred detects more sequences with E-value better than the first false positive compared to HMMER. However, for lower selectivity TPRpred performance is comparable to HMMER.

3.1.4.2 Comparison of the web-servers using STAND family members

To assess the sensitivity of TPRpred in detecting divergent TPR units over Pfam, SMART, and REP, we evaluated the performance of the web-servers using the STAND family test set. Additionally, we also used 53 true negative sequences by selecting arbitrarily from the all-α class of the SCOP database. The hits that were confidently predicted according to the web-servers for the STAND proteins are tabulated in Table 3.2. None of the servers detected false positives from the true negative sequences (data not shown). This shows that all the servers are unbiased to the α-helical proteins which are unrelated.

TPRpred performs significantly better in detecting the TPR units from the

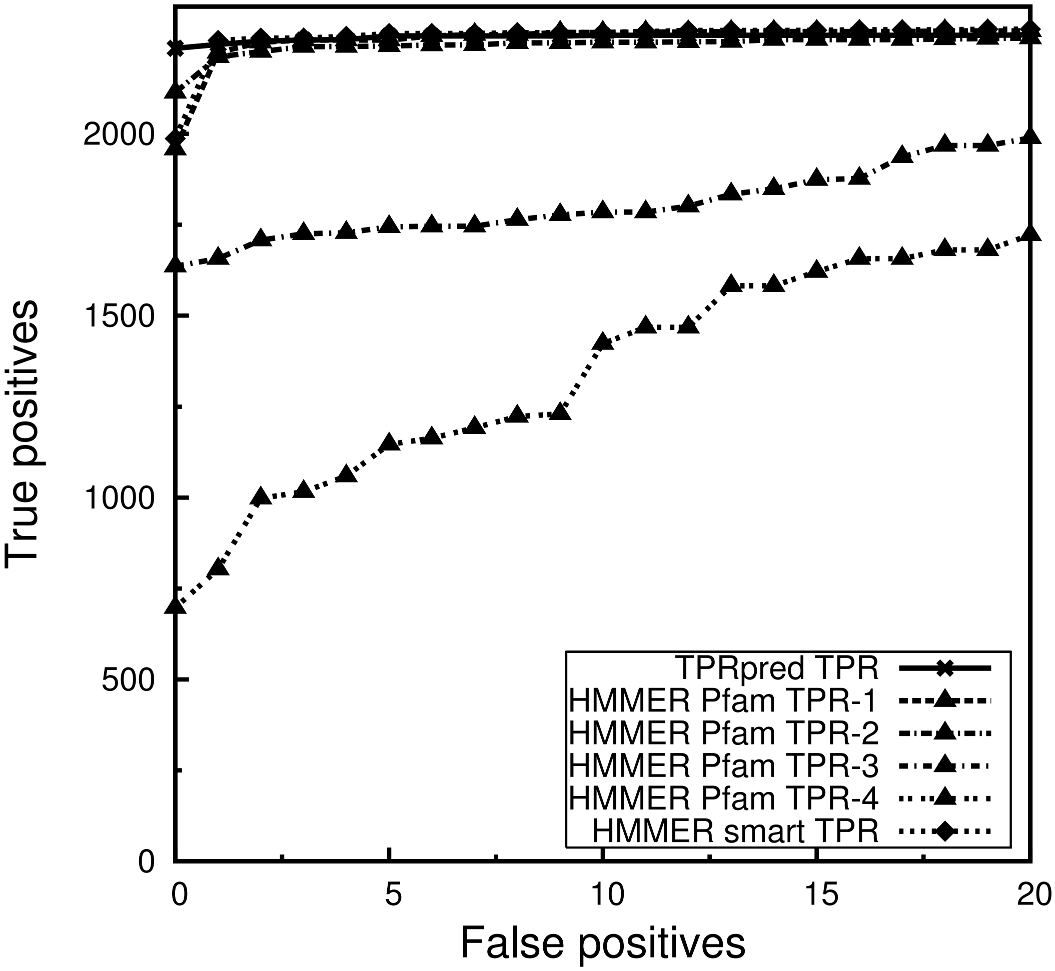


Figure 3.5: **ROC plot comparing the performance of TPRpred and HMMER.** Sensitivity of the methods, measured by the number of true positives detected at varying numbers of false positives.

Table 3.2: **Comparison of the results obtained from the web-servers using a set of 56 STAND family members.** The number of STAND members and individual repeats detected are tabulated.

	TPRpred	**Pfam**	**SMART**	**REP**
Proteins detected (% of total)	48 (85 %)	24 (42 %)	6 (10 %)	5 (8 %)
Individual repeats detected	302	50	30	35

members of the STAND family, although sequences of the STAND family members were explicitly excluded from our TPR profile. For instance, the 8 TPR units present in MalT [115] were detected only by our server. Overall, TPRpred detected twice as many proteins as TPR-containing proteins and over 6 fold more individual repeats as the next best web-server, Pfam. This could be due to the more sensitive Gaussian scoring as well as the score base-line strategy employed by our tool.

3.1.4.3 Comparison of the web-servers using known protein structures

In order to assess the sensitivity of the web-servers in detecting the individual repeat units, we submitted the sequences of the TPR structure set, along with 2 SEL1-like repeat proteins classified under the HCP-like family [SCOP:a.118.18.1], to TPRpred, Pfam, SMART, and REP web-servers. The number of repeats detected confidently for each sequence are tabulated in Table 3.3 and the repeats detected only by TPRpred are shown in Figure 3.6. The TPR structure set contains both proteins that were present in the training databases of the individual methods (Table 3.3, top) and proteins whose structure became available subsequently (Table 3.3, bottom). All servers performed well on the former proteins, although TPRpred stood out with 100 % detected individual repeats over the other servers, which only detected between 70 % and 90 %, but the real difference between servers became visible on the latter proteins. Here, TPRpred recognized all proteins as TPR-containing, whereas the other servers recognized less than half, and TPRpred detected 94 % of individual repeats, whereas the other servers detected only about 48 %.

Table 3.3: **The comparison of the results obtained from the web-servers using known structures.** The actual number of repeats for each entry and the number of repeats detected by various web-servers are tabulated. See also Figure 3.6.

PDB-ID	Name	Type	#¶	TPRpred	Pfam	SMART	REP
Structures used in profile generation by TPRpred							

Continued on next page

Table 3.3 – continued from previous page

PDB-ID	Name	Type	#[¶]	TPRpred	Pfam	SMART	REP
1A17	Phosphatase 5	TPR	3	3	3	3	0
1KT1	Fkbp51	TPR	3	3	2	2	3
1ELR	Hop, TPR2a	TPR	3	3	3	3	3
1IHG	Cyclophilin 40	TPR	3	3	3	3	3
1ELW	Hop, TPR1	TPR	3	3	3	3	3
1HH8	P67phox	TPR	3	3	3	3	3
1FCH[§]	PEX5, human	TPR	7	7	4	4	6
1HXI	PEX5, brucei	TPR	3	3	3	3	3
1KLX	Hcpb	SEL1	3	3	3	3	3
1OUV	Hcpc	1[†]+6[‡]	7	1[†]+6[‡]	1[†]+6[‡]	7[‡]	7[†]
Total			**38**	**38**	**34**	**33**	**27**
% of total				**100 %**	**89 %**	**86 %**	**71 %**
Structures not used in profile generation by TPRpred							
1P5Q	Fkbp52	TPR	3	3	3	3	3
2C2L	CHIP	TPR	3	3	3	3	3
1XNF[§]	Nlpi	TPR	4	4	3	3	3
1W3B[§]	Transferase	TPR	10	10	9	9	9
1TJC[§]	Hydroxylases	TPR	2	2	1	1	0
1HZ4	MalT	TPR	8	8	0	0	0
1NZN	Fis1	TPR	2	1	0	0	0
1YA0	Smg7	TPR	2	2	0	0	0
1ZU2	Tom20, plant	TPR	2	2	0	0	0
1IYG	Rsgi Ruh-001	TPR	2	1	0	0	0
1OM2	Tom20, animal	TPR	1	1	0	0	0
Total			**39**	**37**	**19**	**19**	**18**
% of total				**94 %**	**48 %**	**48 %**	**46 %**

[¶] Actual number of repeats in the structure

[§] Structures shown in Figure 3

[†] TPR

[‡] SEL1-like repeat

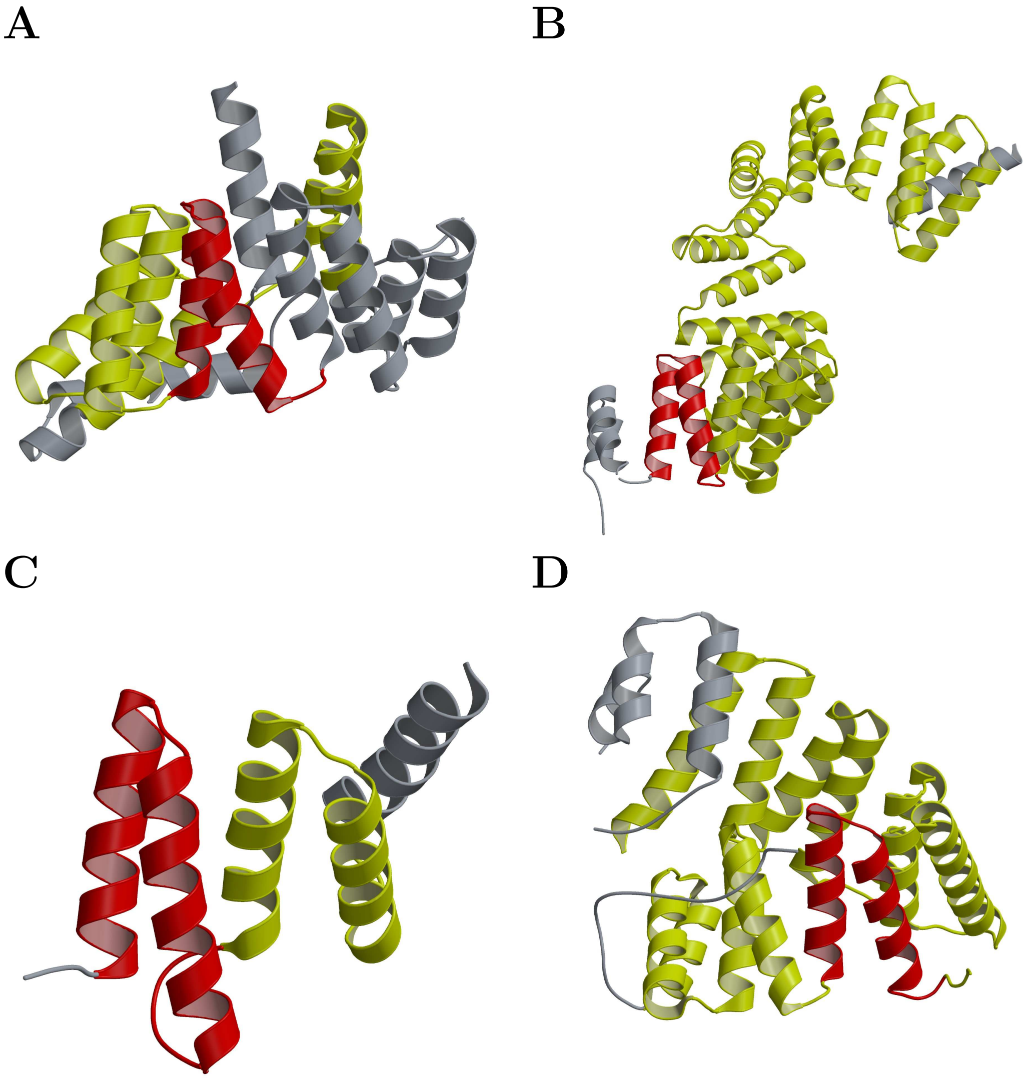


Figure 3.6: **The accuracy of TPRpred in detecting individual repeats.** The TPRs detected only by TPRpred are shown in red, whereas TPRs also detected by the other servers are shown in yellow, and the remaining residues are shown in grey. Structures in which all TPRs are only recognized by TPRpred are omitted. (A) *E. coli* NlpI [PDB:1XNF, chain A]. (B) Human N-acetylglucosamine transferase, TPR domain [PDB:1W3B, chain A]. (C) Peptide-substrate-binding domain of human type I collagen prolyl 4-hydroxylase [PDB:1TJC, chain A]. (D) Human PEX5 [PDB:1FCH, chain A]. The figure was generated using MOLSCRIPT [145] and Raster3D [146].

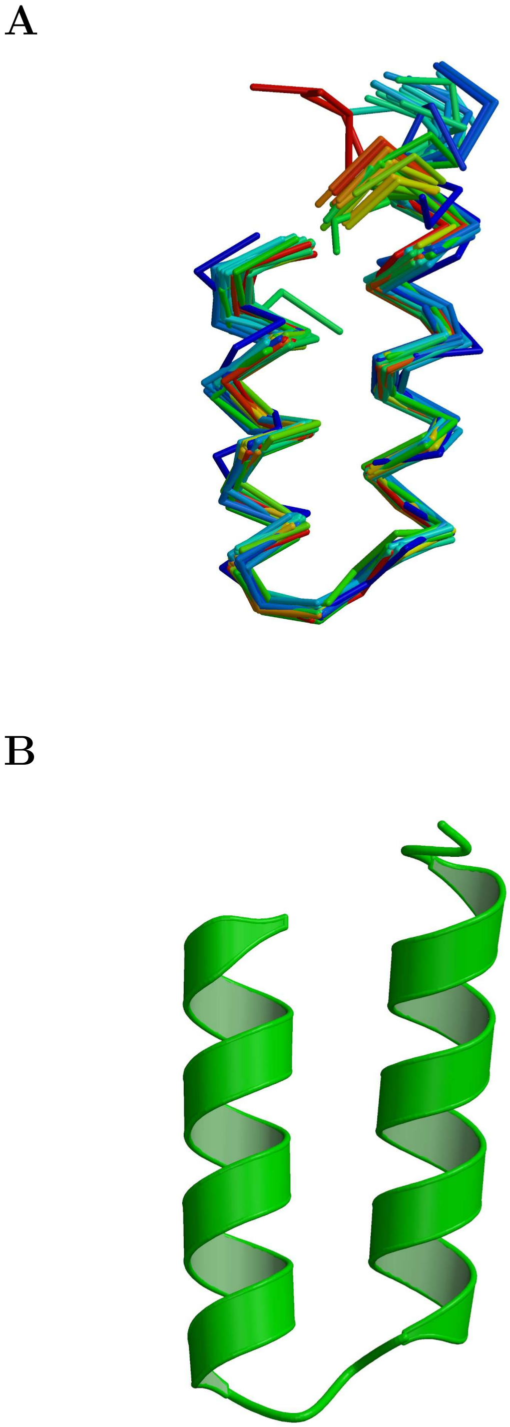


Figure 3.7: **Superposition and computed average TPR unit**. (A) Superposition of Cα traces of 44 individual TPR units from the 12 TPR-containing structures. The individual TPR units are shown in different colours. (B) The average TPR unit derived from the superposed repeat units.

3.1.5 TPR-like hairpins in globular proteins

To detect the TPR-like repeat units in non-TPR-containing structures, we used a representative TPR unit, which we obtained by superimposing 40 individual TPR units from 12 TPR-containing structures (Figure 3.7A), and derived the average TPR unit (Figure 3.7B). The maximum backbone RMSD of 2.5 $\AA$ was observed over 26 residues. The region connecting the two helices are highly conserved compared to the termini among the individual repeat units (Figure 3.7A).

The description of protein structure in the language of side chain contact maps is shown to offer many advantages over more traditional approaches. The supersecondary structural fragments that occurs in non-related proteins are characterized by small number of interactions patterns which can be captured by structural similarity measured by the overlap of the side chain contact maps. Therefore, we employed contact map based structure comparison over more traditional approaches to measure the structural similarity. The contact map based structure comparison of the average TPR unit with non-TPR-containing structures revealed 51 α-hairpins with greater than 80 % structural similarity to the average TPR unit which are tabulated in Table 3.4. The 92 % of these hits belong to the all alpha class of the SCOP classification. This reveals that TPR-like units are less common in other protein classes which have both α and β. This could be attributed to the less occurrence of α-hairpin supersecondary structural motif as compared to $\alpha\beta$-motif and β-hairpin [147]. Similarly, 27 % of the hits were proteins that belong to the ferritin-like superfamily (a.25.1) and 21 % belong to the up-down helical bundle fold. The remaining hits were distributed over different folds such as $\alpha\alpha$-superhelix, spectrin repeat-like, phospholipase CP1, SAM domain-like and many more. It is interesting to note that ferritin-like, up-down helical bundle, SAM domain-like and $\alpha\alpha$-superhelix folds are the members of the super-fold [75, 79].

Structural similarity between two protein fragments could be due to divergent or convergent evolution. In order to select only homologous α-hairpins over analogous ones, we have extracted the primary amino acid sequences corresponding to the α-hairpin regions from each of the hits and ranked them using the profile-to-sequence comparison program TPRpred as shown in Table 3.5. TPRpred uses the sequence information of the homologous repeats in the profile, thereby we

Table 3.4: **Summary of structure-structure comparison.** The best 22 structurally similar $\alpha\alpha$-hairpins to the average TPR units are listed

Nr.	**PDB-ID**	**Res**[*]	**SCOP**[†]	**score**[‡]	**%score**[§]	**Rep**[¶]	**Name**
1	1KHOA	98-130	a.124.1.1	197	94.3	1	Alpha-toxin
2	2E2AA	20-52	a.7.2.1	197	94.3	1	Enzyme IIa
3	1JGCA	98-130	a.25.1.1	196	93.8	2	Bacterioferritin
4	1OCRC	140-172	f.25.1.1	196	93.8	1	cytochrome
5	1JI4A	103-135	a.25.1.1	195	93.3	1	ferritin
6	1LKOA	24-56	a.25.1.1	194	92.8	1	Rubrerythrin
7	1O9DA	66-98	a.118.7.1	194	92.8	2	14-3-3-like
8	1BCFA	22-54	a.25.1.1	192	91.9	1	Bacterioferritin
9	1F4NA	16-48	a.30.1.1	192	91.9	1	ROP
10	1FJGT	33-65	a.7.6.1	192	91.9	1	RPS20
11	1H6GA	397-429	a.24.9.1	192	91.9	2	alpha-catenin
12	1H6GB	397-429	a.24.9.1	192	91.9	2	alpha-catenin
13	1JGCA	22-54	a.25.1.1	192	91.9	2	Bacterioferritin
14	1NFVA	27-59	a.25.1.1	192	91.9	1	Bacterioferritin
15	1EEXG	83-115	a.23.2.1	191	91.4	1	Diol dehydratase
16	1JMSA	169-201	a.60.6.1	191	91.4	1	deoxynucleotidyl
17	1LF6A	452-484	a.102.1.5	191	91.4	1	glucoamylase
18	1QGHA	110-142	a.25.1.1	191	91.4	1	ferritin
19	1CA1	98-130	a.124.1.1	189	90.4	1	Alpha-toxin
20	1FPOA	125-157	a.23.1.1	189	90.4	1	HSC20
21	1F4NB	16-48	a.30.1.1	188	90.0	1	ROP
22	1OM2A	20-52	a.23.4.1	188	90.0	1	Tom20

[*] Begining and end residue number

[†] SCOP classification identifier

[‡] Structural similarity between contact maps

[§] % Similarity score to query structure

[¶] Number of repeats detected

Table 3.5: **Summary of sequence based ranking.** The amino acid sequences corresponds to the α-hairpins are ranked using TPRpred.

Nr.	PDB-ID	Score	P-value	E-value	Prob[§]	Name
1	1OM2A	21.2	4.1E-07	2.1E-05	20.89	Tom20
2	1FJGT	21.1	4.5E-07	2.3E-05	19.84	RPS20
3	1KHOA	16.3	1.2E-05	6.3E-04	2.22	Phospholipase C
4	1CA1	16.1	1.4E-05	7.1E-04	2.03	Alpha-toxin
5	1O9DA	12.5	1.7E-04	8.8E-03	0.30	14-3-3-like
6	1FPOA	12.4	1.9E-04	9.6E-03	0.28	HSC20
7	1LF6A	12.3	2.0E-04	1.0E-02	0.26	glucoamylase
8	1AFRA	10.7	6.0E-04	3.1E-02	0.11	Desaturase
9	1QJBA	10.6	6.4E-04	3.3E-02	0.10	14-3-3
10	1AH7	10.5	6.7E-04	3.4E-02	0.10	Phospholipase C

[§] Probability for being a TPR repeating unit

eliminated all the analogous α-hairpins. The 1st hit was from the mitochondrial translocase of outer membrane complex (1OM2A, Tom20). Tom20 is indeed an example for one repeat TPR-containing proteins [126], which demonstrates that our approach to identify TPR-like α-hairpins in non-TPR-containing proteins appear to be correct. The 2nd hit was from the Ribosomal Protein S20 (1FJGT, RPS20). Ribosomal proteins are most ancient molecules and many of the modern non-ribosomal proteins contain fragments from ribosomes. Therefore, ribosomes might have played a major role in early protein evolution [12]. The 3rd hit was α-toxin of the phospholipase C family (PLC). The 4th and 5th hits are less interesting since the 4th hit is a homolog of α-toxin (3rd hit), while the 5th hit belong to the same fold as TPR. Additionally, we considered the 6th (heat shock protein 20, HSC) and 7th (bacterial glucoamylase, BGA). Overall, we selected five α-hairpins for experimental studies, and whose structures are shown in Figures 3.8, and 3.9.

3.1.6 Design of TPR-like domains

The central theme of engineering novel proteins is the ability to design specific amino acid sequences that fold into desired structures. Protein design is also a good method to assess our understanding of sequence-structure and structure-

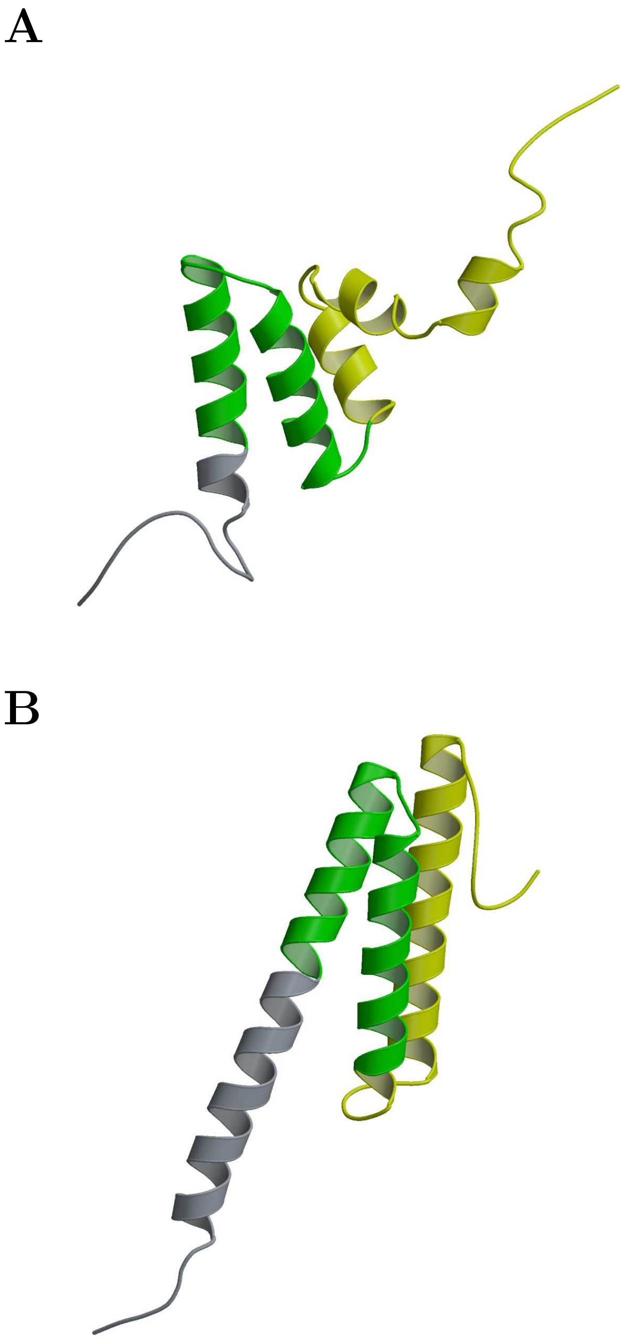


Figure 3.8: **Proteins selected for experimental study with natural "solvating" or "stop" helix**. The two representative structures of the total five proteins selected for further study categorised based on the solvating helix. TPR-like α-hairpins are shown in green, the "solvating" helices are shown in yellow and the remaining residues are shown in grey. The PDB codes are: (A) Tom20 (1OM2A), (B) RPS20 (1FJGT)

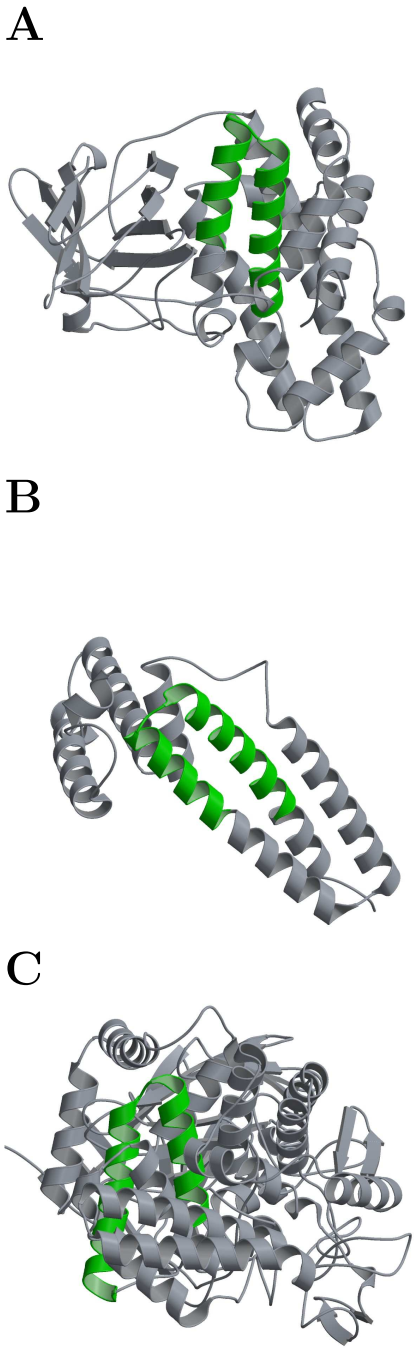


Figure 3.9: **Proteins selected for experimental study with no natural "solvating" or "stop" helix**. The three representative structures of the total five proteins selected for further study that have no natural solvating helix. TPR-like α-hairpins are shown in green and the remaining residues are shown in grey. The PDB codes are: (A) PLC (1KHOA), (B) HSC (1FPOA) and (C) BGA (1LF6A).

function relationships. Many proteins aggregate when removed from their physiological context. The aggregation of proteins will be much more in the proteins where amino acid sequences are mutated from their natural sequences.

TPR domains are of interest with respect to their folding, modular architecture, and range of binding specificities. Regan *et al.* designed idealized TPR repeating unit which allowed them to show how the amino acid sequence specifies fold and function [123, 124, 125]. They used a consensus-based TPR design to engineer novel proteins by arraying varing numbers of an idealized TPR repeating unit. Further, they constructed three separate proteins with one, two, and three copies of the consensus TPR repeating unit, thereby not only showing that the resultant statistically designed proteins were stable and have native-like properties, but also correctly formed the desired TPR fold as demostrated by both NMR and X-ray crystallography. By this way, they pioneered in making perfectly repetitive TPR-containing proteins, which are identical both at the sequence and structural level that are not exist in the nature. We adopted similar designing strategy however by using the α-hairpin protein fragments in constrast to the consensus-based TPR unit. From each of the five α-hairpins three seperate proteins were constructed with one, two and three copies of the α-hairpins. Further, we inserted three extra features into each of designed proteins (Figure 3.10):

- The sequence Gly-Asn-Ser was added at the N-terminus of each protein to provide a potential N-capping, helix-stabilizing sequence. Statistically, Gly, Asn, and Ser have the highest propensities to occur at the N'', N', and N cap positions in α-helices [148, 149, 150, 151].
- The C-terminal four residues of each hairpin which constitute a loop connecting the juxtaposed repeat were substituted by Asp-Pro-Asn-Asn to maintain the angle between juxtaposed repeat units. Statistically, Asp, Pro, Asn and Asn have the highest propensities to occur at these positions. This 4 residues mutation were not made in one repeat proteins with natural ”solvating” helix namely Tom20-1 and RPS20-1 proteins.
- An additional artificial solvating helix was added after the final helix of the repeat unit for those α-hairpin which could not be assigned the natural solvating helix after examining their parent protein structures.

All of the information needed to specify a protein's 3D structure is contained within its amino-acid sequence. More similar the sequences are to the known TPRs, the higher is the probability that they would adopt a TPR fold. Therefore, we have explored the sequence space for these α-hairpins by obtaining homologs by BLAST searches against the non-redundant database and selecting the best sequence using TPRpred.

3.1.6.1 Design of Tom20-1, Tom20-2 and Tom20-3

Mitochondria are generally accepted to have descended from a eubacterium that was engulfed by an archaebacterial host cell [152, 153]. During the evolution of this endosymbiotic relationship, the vast majority of organellar genes were transferred to the nucleus, necessitating an efficient system to import nuclear-encoded mitochondrial proteins into the organelle [153]. All extant mitochondria possess this protein import machinery, consisting of the translocase of the outer mitochondrial membrane (TOM) complex, and two inner mitochondrial membrane complexes [154]. The TOM complex consists of two functionally defined groups of proteins that either form the 'general import pore' or act as receptors that facilitate delivery of the precursor protein to the pore [155, 156].

The receptor protein Tom20 from *Rattus norvegicus* (rat) is anchored in the mitochondrial outer membrane by an N-terminal hydrophobic transmembrane segment, and the C-terminal domain, which functions as a receptor for mitochondrial precursor proteins, that is exposed to the cytosol. This cytosolic domain of rat Tom20 consists of two acidic regions, a Q-rich region containing glutamine and hydrophobic residues, and a single TPR unit [126]. This protein was ranked first in our final sequence based ranking of the α-hairpins that were detected by structure-structure comparisons. This hit served as our positive control for the approach we took in finding the TPR-like repeating units in non-TPR-containing proteins, because of its clear nature of possessing single TPR unit.

We extracted a set of 18 homologs of Tom20 protein sequence conservatively with BLAST (E-value cutoff of 10^{-3}) from the non-redundant database at NCBI (`http://www.ncbi.nlm.nih.gov`), using the cystosolic domain of Tom20 [GI:6980855] as a query sequence. Each of these sequences were feeded to TPRpred program to find out their sequence fitness to the known TPR units. Results from these comparisons are tabulated in Table 3.6, which show that Tom20 from

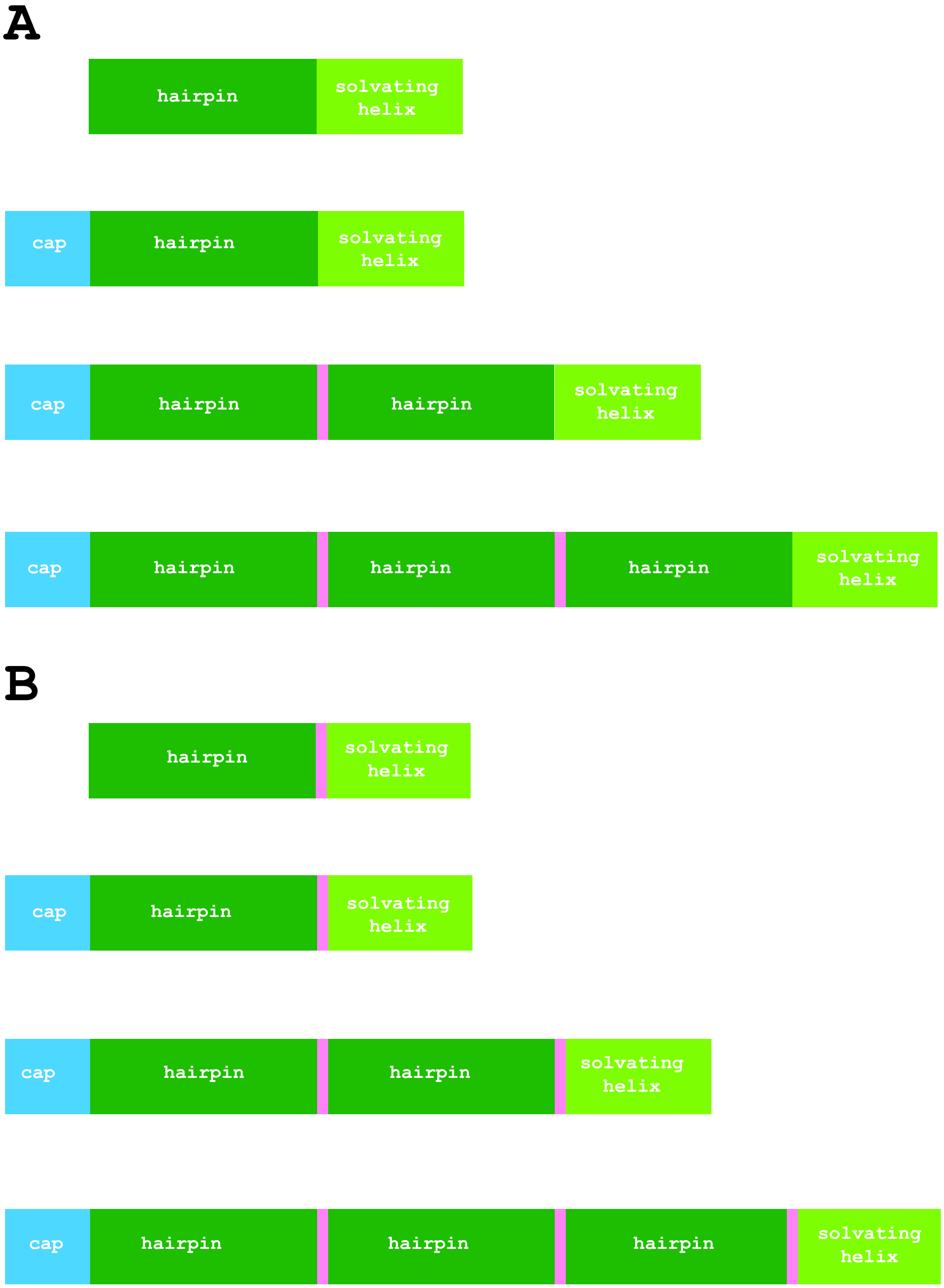


Figure 3.10: **Schematic illustration of the designed proteins.** The colour codes indicates N cap in cyan, α-hairpin in green, turn in magenta and solvating helix in yellowish green. (A) Tom20 and RPS20 derived proteins. (B) PLC, HSC and BGA derived proteins.

Xenopus laevis was ranked best. However, we selected the second hit that was our query sequence, because structure of this protein is known, and moreover, the differences between these two hits is not much stastically.

To construct the plasmid containing Tom20-1 gene, we used the cDNA of hippocampal cells from *Rattus norvegicus* as a template with appropriate primers by PCR. Finally, we constructed the genes encoding the two and three identical copies of the repeats with non-overlapping nucleotide sequences that are produced by assigning the first and second most favoured codon usage for second and third repeats of amino acid sequences respectively. The DNA and protein sequences are shown in Figure 3.11.

Table 3.6: **Homologs of Tom20 proteins. The TPR-like repeats from the set of 18 Tom20 proteins are tabulated in the order from best to worst based on their single repeat P-values for being TPR repeating unit. The idealized TPR unit[§] sequence is given for easy comparison.**

GI number	Sequence
Idealized TPR[§]	`AEAWYNLGNAYYKQGDYDEAIEYYQKALELDPNN`
27371034	`FLEEIQLGEELLAQGDFEKGVDHLTNAIAICGQP`
6980855[¶][†]	`FLEEIQLGEELLAQGDYEKGVDHLTNAIAVCGQP`
37546318	`FFEEIQLGEELLAQGEYEKGVDHLTNAIAVCGQP`
23097350	`FLEEIQLGEELLAQGDYEKGVDHLTNAIAVCGQP`
1838935	`FLEEIQLGEELLAQGDYEKGVDHLTNAIAVCGQP`
13324686	`FLEEIQLGEELLAQGDYEKGVDHLTNAIAVCGQP`
15741057	`FLEEIQLGEELLAQGEYEKGVDHLTNAIAVCGQP`
37589685	`FLDEIQLGEELLAQGDYEQGVDHLTNAIAVCGQP`
37541284	`FLEETQLGEELLAQGKYEKGVDHLTNAIAVCGQP`
7657257	`FLEEIQLGEELLAQGEYEKGVDHLTNAIAVCGQP`
38086424	`FLEEIQLGEELLAQGDYEKGVDHLTNATAVCGQP`
24645735	`FMTQIHKGETLITNGDVEAGVEHLINAILVCGQP`
28574622	`FLQEIQLGETLIARGDFESGVEHLANAIVVCGQP`
19909174	`FLQQIQQGETALSMGSLDEVVNHFAIAVSICCQP`
37181652	`GELWLSRGEHRMGIQHLGNALLVCEQPRELLKVF`

Continued on next page

Table 3.6 – continued from previous page

GI number	Sequence
31239595	`FLQEIQTGEALISAGDIENGVEHLANAIIVCGQP`
39592238	`AVMLCGESQQLLSIFQQTLSEEQFRAVVQQLPST`
17560096	`DEGAVHIANAVMLCGESQQLLSIFQQTLSEDQFR`

§ Consensus sequence for idealized TPR from Regan *et al.*
† Finally selected α-hairpin
¶ Query sequence hairpin

3.1.6.2 Design of RPS20-1, RPS20-2 and RPS20-3

Protein synthesis is coordinated by the ribosomes and is fundamentally same in all cells. The structure of the bacterial ribosome has been studied for many years, and the two subunits together form a 70S particle that has an approximate molecular mass of 2.3 MDa. In general the small, 30S subunit (S) contains a 16S rRNA and 21 proteins; the large, 50S subunit (L) contains a 23S rRNA, a 5S rRNA and 34 proteins, however the exact number of proteins varies between species. An important role of ribosomal proteins is to direct the folding and stabilize the tertiary structure of rRNA. Consistent with this, the majority of the proteins appear to have multiple RNA-binding sites and probably interact with several regions of rRNA. Several of these ribosomal-protein folds appeared to be unique, when they were first discovered, but were subsequently found to be present in other proteins [12]. An RNA world is widely accepted as a probable stage in the early evolution of life. The ribosomes are the relics of the RNA world and have many features that are remnants of the RNA world. Its appearance was obviously a crucial step in early protein evolution. Therefore, protein evolution studies based on ribosomal proteins play a significant role.

Protein domains are supposed to have arisen divergently from a basic set. There is a growing body of evidence suggesting that this basic set arose by duplication, mutation and shuffling of shorter fragments. Since proteins today are often a mosaic of homologous and nonhomologous domains, the domains themselves too may be mosaics of homologous and nonhomologous fragments. The fragments presumably evolved in the context of RNA-based replication and catalysis and would also have converged towards particularly foldable and versatile structural

>**Tom20-1 gene**
GGAATTCCATATGGGCAACAGCTTCCTTGAAGAGATACAGCTTGGTGAAGAGTTATTAGC
ACAAGGTGACTATGAGAAGGGTGTGGACCACCTGACAAATGCAATCGCTGTGTGTGGACA
GCCTCAGCAGTTGCTGCAAGTGTTACAACAGACTCTTCCACCACCAGTGTTCCAGATGCT
TCTGACCAAGCTTCCAACCATTAGTCAGAGAATTGTCAGTGCTCAGAGCTTGGCTGAGGA
TGATGTGGAATGACTCGAGCGG

>**Tom20-1 protein**
MGNSFLEEIQLGEELLAQGDYEKGVDHLTNAIAVCGQPQQLLQVLQQTLPPPVFQMLLTK
LPTISQRIVSAQSLAEDDVE

>**Tom20-2 gene**
GGAATTCCATATGGGCAACAGCTTTCTGGAAGAAATTCAGCTGGGCGAAGAACTGCTGGC
GCAGGGCGATTATGAAAAAGGCGTGGATCATCTGACCAACGCGATTGCGGTGGATCCGAA
CAACTTCCTTGAAGAGATACAGCTTGGTGAAGAGTTATTAGCACAAGGTGACTATGAGAA
GGGTGTGGACCACCTGACAAATGCAATCGCTGTGTGTGGACAGCCTCAGCAGTTGCTGCA
AGTGTTACAACAGACTCTTCCACCACCAGTGTTCCAGATGCTTCTGACCAAGCTTCCAAC
CATTAGTCAGAGAATTGTCAGTGCTCAGAGCTTGGCTGAGGATGATGTGGAATGACTCGA
GCGG

>**Tom20-2 protein**
MGNSFLEEIQLGEELLAQGDYEKGVDHLTNAIAVDPNNFLEEIQLGEELLAQGDYEKGVD
HLTNAIAVCGQPQQLLQVLQQTLPPPVFQMLLTKLPTISQRIVSAQSLAEDDVE

>**Tom20-3 gene**
GGAATTCCATATGGGTAATTCGTTCTTGGAGGAGATCCAATTGGGTGAGGAGTTGTTGGC
CCAAGGTGACTACGAGAAGGGTGTTGACCACTTGACGAATGCCATCGCCGTTGACCCAAA
TAATTTTCTGGAAGAAATTCAGCTGGGCGAAGAACTGCTGGCGCAGGGCGATTATGAAAA
AGGCGTGGATCATCTGACCAACGCGATTGCGGTGGATCCGAACAACTTCCTTGAAGAGAT
ACAGCTTGGTGAAGAGTTATTAGCACAAGGTGACTATGAGAAGGGTGTGGACCACCTGAC
AAATGCAATCGCTGTGTGTGGACAGCCTCAGCAGTTGCTGCAAGTGTTACAACAGACTCT
TCCACCACCAGTGTTCCAGATGCTTCTGACCAAGCTTCCAACCATTAGTCAGAGAATTGT
CAGTGCTCAGAGCTTGGCTGAGGATGATGTGGAATGACTCGAGCGG

>**Tom20-3 protein**
MGNSFLEEIQLGEELLAQGDYEKGVDHLTNAIAVDPNNFLEEIQLGEELLAQGDYEKGVD
HLTNAIAVDPNNFLEEIQLGEELLAQGDYEKGVDHLTNAIAVCGQPQQLLQVLQQTLPPP
VFQMLLTKLPTISQRIVSAQSLAEDDVE

Figure 3.11: **The DNA and protein sequences of Tom20 derived α-hairpin.** The colour codes indicates N cap in cyan, α-hairpin in green, turn in magenta and solvating helix sequences in yellow. For DNA sequences the NdeI and XhoI restriction sites are displayed with an underline. Additional nucleotides (black) required by the restriction enzymes for effective digestion are also displayed at both the termini of the genes

solutions (the supersecondary structures). Their combination, often through repetition of the same element, was the driving force behind the evolution of folded domains [79]. Therefore, our discovery of the α-hairpin from the ribosomal protein S20 (RPS20) that resemble the repeating units of TPRs play a significant role.

In order to find out the α-hairpin from the homologs of RPS20 which is similar to the known repeating units of TPRs at the sequence level, we extracted a set of 7 homologs conservatively with BLAST (E-value cutoff of 10^{-3}) from the non-redundant database, using 30S ribosomal protein S20 (RPS20) [GI:10835604] from *Thermus Thermophilus* as a query sequence. Each of these sequences were feeded to TPRpred program to find out their sequence fitness to the known TPR units. Results from these comparisons are tabulated in Table 3.7. We selected our query sequence as a best hairpin.

To construct the plasmid containing RPS20-1 gene, we used the genomic DNA from *Thermus Thermophilus* as a template with appropriate primers by PCR. Finally, we constructed the genes encoding the two and three identical copies of the repeats with non-overlapping nucleotide sequences that are produced by assigning the first and second most favoured codon usage for second and third repeats of amino acid sequences respectively. The DNA and protein sequences are shown in Figure 3.12.

Table 3.7: **Homologs of ribosomal protein S20 (RPS20). The α-hairpins from the set of 6 RPS20 proteins are listed in the order from best to worst based on their single repeat P-values for being TPR repeating unit. The idealized TPR unit§ sequence is given for easy comparison.**

GI number	**Sequence**
Idealized TPR§	`AEAWYNLGNAYYKQGDYDEAIEYYQKALELDPNN`
10835604¶†	`IKTLSKKAVQLAQEGKAEEALKIMRKAESLIDKA`
34811548	`IKTLSKKAVQLAQEGKAEEALKIMRKAESLIDKA`
14278549	`IKTLSKKAIQLAQEGKAEEALKIMRKAESLIDKA`
20095227	`IKTLSKKAIQLAQEGKAEEALKIMRKAESLIDKA`
1350957	`LKTIEKRCINMIKAGKKDEAIEFFKFVAKKLDTA`

Continued on next page

Table 3.7 – continued from previous page

GI number	Sequence
15594578	`LKTIEKRCINMIKAGKKDEAIEFFKFVAKKLDTA`

§ Consensus sequence for idealized TPR from Regan *et al.*
† Finally selected α-hairpin
¶ Query sequence hairpin

3.1.6.3 Design of PLC-1, PLC-2 and PLC-3

Enzymes are under high selective pressure therefore they play less significant role in understanding the protein evolution. The α-toxin from *Clostridium perfringens* was the first bacterial protein shown to have both enzymatic and toxic properties. The phospholipase C (PLC) is a zinc metalloenzyme, composed of 370 residues that exhibits hemolytic, necrotic, vascular permeabilization, and platelet aggregating properties. It is predominantly associated with the disease gas gangrene in humans. The enzyme consists of two domains. The N-terminal domain (residues 1-246) is α-helical and contains the active site, which is indicated by the presence of either two or three metal ions. The C-terminal domain, which is known to be essential for toxicity, is an eight-stranded β-sandwich [157]. An α-hairpin containing residues 98-130 in the N-terminal domain of PLC resembles the repeating unit of TPR.

Table 3.8: **Homologs of phospholipase C (PLC) proteins. The α-hairpins from the set of 32 PLC proteins are listed in the order from best to worst based on their single repeat P-values for being TPR repeating unit. The idealized TPR unit§ sequence is given for easy comparison.**

GI number	Sequence
Idealized TPR§	`AEAWYNLGNAYYKQGDYDEAIEYYQKALELDPNN`
23821941†	`VRKFTALARNEWEKGNYEKATWYFGQAMHYFGDL`
21730473	`IRKFSALARYEWKRGNYKQATFYLGEAMHYFGDA`
29373300	`GTKYFNISIEEYQDGNFEKAFYNLGLAIHYYTDI`

Continued on next page

Table 3.8 – continued from previous page

GI number	Sequence
2126777	`PDNLSNKGFVYTKIKKNDRFVHVIGTHLQAEDSM`
30145480	`IRKFSALARYEWQRGNYKQATFYLGEAMHYFGDI`
4929954	`IRKFSALARYEWQRGNYKQATFYLGEAMHYFGDI`
6137451	`IRKFSALARYEWQRGNYKQATFYLGEAMHYFGDI`
23821940	`LRKLFALAKDEWKKGNYEQATWLLGQGLHYFGDF`
1255553	`IRKFSALARYEWQRGNYKQATFYLGEAMHYFGDI`
1255555	`IRKFSALARYEWQRGNYKQATFYLGEAMHYFGDI`
40555	`IRKFSALARYEWQRGNYKQATFYLGEAMHYFGDI`
1255557	`IRKFSALARYEWQRGNYKQATFYLGEAMHYFGDI`
18309018	`IRKFSALARYEWQRGNYKQATFYLGEAMHYFGDI`
477909	`IRKFSALARYEWQRGNYKQATFYLGEAMHYFGDI`
1741875	`IRKFSALARYEWQRGNYKQATFYLGEAMHYFGDI`
896252	`IRKFSALARYEWQRGNYKQATFYLGEAMHYFGDI`
896256	`IRKFSALARYEWQRGNYKQATFYLGEAMHYFGDI`
896258	`IRKFSALARYEWQRGNYKQATFYLGEAMHYFGDI`
1255545	`IRKFSALARYEWQRGNYKQATFYLGEAMHYFGDI`
1255549	`IRKFSALARYEWQRGNYKQATFYLGEAMHYFGDI`
80525	`IRKFSALARYEWQRGNYKQATFYLGEAMHYFGDI`
98620	`IRKFSALARYEWQRGNYKQATFYLGEAMHYFGDI`
29373282	`GAKYFNQSVTDYREGKFDTAFYKLGLAIHYYTDI`
29373304	`GAKYFNQSVADYREGKFDTAFYKLGLAIHYYTDI`
29373294	`GAKYFNQSVADYREGKFDTAFYKLGLAIHYYTDI`
29373296	`GAKYFNQSVADYREGKFDTAFYKLGLAIHYYTDI`
29373288	`GAKYFNQSVADYREGKFDTAFYKLGLAIHYYTDI`
15004851	`PYHSANLIAVLSTHSQYEQFVQDHQTSYALNSTD`
21730473¶	`IRKFSALARYEWKRGNYKQATFYLGEAMHYFGDA`
282501	`GAKYFNQSVTDYREGKFDTAFYKLGLAIHYYTDI`
16802251	`GAKYFNQSVTDYREGKFDTAFYKLGLAIHYYTDI`
80580	`GAKYFNQSVTDYREGKFDTAFYKLGLAIHYYTDI`
2382197	`IRKFSALARYEWKRGNYKQATFYLGEAMHYFGDA`

§ Consensus sequence for idealized TPR from Regan *et al.*

† Finally selected α-hairpin

¶ Query sequence hairpin

```
>RPS20-1 gene
GGAATTCCATATGGGCAACAGCATCAAGACCCTCAGCAAGAAGGCCGTCCAGCTGGCCCA
GGAGGGCAAGGCGGAAGAGGCCCTGAAGATCATGCGCAAGGCCGAAAGCCTCATTGACAA
GGCGGCGAAGGGCTCCACCCTGCACAAGAACGCCGCCGCCCGCAGGAAGTCCCGGCTGAT
GCGCAAGGTCCGTCAGCTGCTCGAGGCCGCGGGGGCGCCCCTCATTGGCGGCGGCCTCAG
CGCCTAAAAGCTTGGG
```

```
>RPS20-1 protein
MGNSIKTLSKKAVQLAQEGKAEEALKIMRKAESLIDKAAKGSTLHKNAAARRKSRLMRKV
RQLLEAAGAPLIGGGLSA
```

```
>RPS20-2 gene
GGAATTCCATATGGGTAATTCGATCAAGACGTTGTCGAAGAAGGCCATCCAATTGGCCCA
AGAGGGTAAGGCCGAGGAGGCCTTGAAGATCATGCGTAAGGCCGAGTCGTTGGACCCAAA
TAACATCAAGACCCTCAGCAAGAAGGCCGTCCAGCTGGCCCAGGAGGGCAAGGCGGAAGA
GGCCCTGAAGATCATGCGCAAGGCCGAAAGCCTCATTGACAAGGCGGCGAAGGGCTCCAC
CCTGCACAAGAACGCCGCCGCCCGCAGGAAGTCCCGGCTGATGCGCAAGGTCCGTCAGCT
GCTCGAGGCCGCGGGGGCGCCCCTCATTGGCGGCGGCCTCAGCGCCTAAAAGCTTGGG
```

```
>RPS20-2 protein
MGNSIKTLSKKAIQLAQEGKAEEALKIMRKAESLDPNNIKTLSKKAVQLAQEGKAEEALK
IMRKAESLIDKAAKGSTLHKNAAARRKSRLMRKVRQLLEAAGAPLIGGGLSA
```

```
>RPS20-3 gene
GGAATTCCATATGGGGAACTCCATTAAGACTTTATCCAAGAAGGCAATTCAGTTAGCACA
GGAAGGGAAGGCAGAAGAAGCATTAAAGATTATGCGGAAGGCAGAATCCTTAGACCCTAA
CAACATCAAGACGTTGTCGAAGAAGGCCATCCAATTGGCCCAAGAGGGTAAGGCCGAGGA
GGCCTTGAAGATCATGCGTAAGGCCGAGTCGTTGGACCCAAATAACATCAAGACCCTCAG
CAAGAAGGCCGTCCAGCTGGCCCAGGAGGGCAAGGCGGAAGAGGCCCTGAAGATCATGCG
CAAGGCCGAAAGCCTCATTGACAAGGCGGCGAAGGGCTCCACCCTGCACAAGAACGCCGC
CGCCCGCAGGAAGTCCCGGCTGATGCGCAAGGTCCGTCAGCTGCTCGAGGCCGCGGGGGC
GCCCCTCATTGGCGGCGGCCTCAGCGCCTAAAAGCTTGGG
```

```
>RPS20-3 protein
MGNSIKTLSKKAIQLAQEGKAEEALKIMRKAESLDPNNIKTLSKKAIQLAQEGKAEEALK
IMRKAESLDPNNIKTLSKKAVQLAQEGKAEEALKIMRKAESLIDKAAKGSTLHKNAAARR
KSRLMRKVRQLLEAAGAPLIGGGLSA
```

Figure 3.12: **The DNA and protein sequences of RPS20 derived α-hairpin.** The colour codes indicates N cap in cyan, α-hairpin in green, turn in magenta and solvating helix sequences in yellow. For DNA sequences the NdeI and HindIII restriction sites are displayed with an underline. Additional nucleotides (black) required by the restriction enzymes for effective digestion are also displayed at both the termini of the genes

We extracted a set of 32 homologs of PLC conservatively with BLAST (E-value cutoff of 10^{-3}) against the non-redundant database, using the full length sequence of PLC [GI:21730473] as a query sequence. Each of these sequences were feeded to TPRpred program to find out their sequence fitness to the known TPR units. Results from these comparisons are tabulated in Table 3.8, which show that PLC from *Clostridium novyi* was ranked best.

Finally, PLC α-hairpin amino acid sequences with necessary features are converted to a DNA sequences according to the most favoured codon usage in *Escherichia coli* using a Perl script. Different nucleotide sequence that codes for more than one copy of the repeat are produced by assigning the second or third most favoured codon usage for the given amino acid sequence. The DNA and protein sequences are shown in Figure 3.13.

3.1.6.4 Design of HSC-1, HSC-2 and HSC-3

The heat shock co-chaperone (HSC) from *Escherichia coli* is 20 Kda J-type co-chaperone protein that regulates the ATPase activity and peptide-binding activity of HSC66, a HSC70-class molecular chaperone. HSC from *Escherichia coli* consists of two distinct domains, the N-terminal J-domain containing residues 1-75 connected by a short loop to a C-terminal domain containing residues 84-171 [158]. The α-hairpin containing residues 125-157 in the C-terminal domain of HSC resembles the repeating unit of TPR.

Table 3.9: **Homologs of heat shock co-chaperone (HSC) proteins. The α-hairpins from the set of 24 HSC proteins are listed in the order from best to worst based on their single repeat P-values for being TPR repeating unit. The idealized TPR unit[§] sequence is given for easy comparison.**

GI number	Sequence
Idealized TPR[§]	`AEAWYNLGNAYYKQGDYDEAIEYYQKALELDPNN`
37527160[†]	`IKTRSQLMVQQLDEQQWEQAADTVRKLRFLDKLQ`
40745443	`PELLMEVMDVQEAIEEVGEGQEAVEKIAVMKKEN`
24373820	`ETTTVKDTAFLMQQMEWREALEDIRESIDHQAII`

Continued on next page

Table 3.9 – continued from previous page

GI number	Sequence
12084584¶	FDTRHQLMVEQLDNETWDAAADTCRKLRFLDKLR
33112677	AKTAQEKAQNLITSTELNKAYSTLKDALKRAEYM
23956362	QKEFTDNINSAFEQGDFEKAKELLTKMRYFSNIE
22998537	TKGWFMSLHACWSCKGPVEASAFCPTCGAIQPPQ
31228753	QKSEHEKAIALEWSSLVNKAYKTLSKSIERGKYL
9758421	LKQWSDSFVEAFESQKFDDAVKCIQRMTYYERAC
30681429	LKQWSDSFVEAFESQKFDDAVKCIQRMTYYERAC
21672835	IDNYEKIIEIKFNEKKWDDIIKLITKLLFFKKIQ
15599006	AELEREFAACWDDAQRREEAERLVRRMQFLDKLA
15832647	FDTRHQLMVEQLDNETWDAAADTVRKLRFLDKLR
24113856	FDTRHQLMVEQLDNEAWDAAADTVRKLRFLDKLR
16123084	AQLWVQAADTVRKLRFLDKLQQQVEQLEERLFDD
29250090	KNAKKLSEMRHALTKRFSDCSELFGKALAEKDSN
33152209	LAALQQSADVNEALQILKDPIARATAIIEINTGI
40063436	TSSENEKIQSMIKSTQTNDAFQTLKSPIKRAKYI
33112306	EQLNEGFAACWADPRRRDEAERLARRMQFLDKLF
19073989	RARYLSNAKKLSVDKKFLEDVLEYEEAISNISSD
23105217	EQLNEGFAACWADPRRRDEAERLARRMQFLDKLF
19115581	DSPEKLLQLSQENQGRKVQEINEIRKAMESSNWD
34872990	QKEFTDNINRAFEQGDFEKAKELLTKMRYFSNIE

§ Consensus sequence for idealized TPR from Regan *et al.*

† Finally selected α-hairpin

¶ Query sequence hairpin

We extracted a set of 24 homologs of HSC conservatively with BLAST (E-value cutoff of 10^{-3}) against the non-redundant database, using the full length sequence of HSC [GI:12084584] as a query sequence. Each of these sequences were feeded to TPRpred program to find out their sequence fitness to the known TPR units. Results from these comparisons are tabulated in Table 3.9, which shows that co-chaperone HSC from *Photorhabdus luminescens* was ranked best.

Finally, HSC α-hairpin amino acid sequences with necessary features were converted to a DNA sequences according to the most favoured codon usage in *Escherichia coli* using a Perl script. Different nucleotide sequence that codes for

>**PLC-1 gene**
GGAATTCCATATGGGCAACAGCGTGCGCAAATTTACCGCGCTGGCGCGCAACGAATGGGA
AAAAGGCAACTATGAAAAAGCGACCTGGTATTTTGGCCAGGCGATGCATTATGATCCGAA
CAACGCGGAAGCGAAACAGGACTTGGGTAATGCCAAGCAAAAGCAAGGTTGACTCGAGCG
G

>**PLC-1 protein**
MGNSVRKFTALARNEWEKGNYEKATWYFGQAMHYDPNNAEAKQDLGNAKQKQG

>**PLC-2 gene**
GGAATTCCATATGGGTAATTCGGTTCGTAAGTTCACGGCCTTGGCCCGTAATGAGTGGGA
GAAGGGTAATTACGAGAAGGCCACGTGGTACTTCGGTCAAGCCATGCACTACGACCCAAA
TAATGTGCGCAAATTTACCGCGCTGGCGCGCAACGAATGGGAAAAAGGCAACTATGAAAA
AGCGACCTGGTATTTTGGCCAGGCGATGCATTATGATCCGAACAACGCGGAAGCGAAACA
GGACTTGGGTAATGCCAAGCAAAAGCAAGGTTGACTCGAGCGG

>**PLC-2 protein**
MGNSVRKFTALARNEWEKGNYEKATWYFGQAMHYDPNNVRKFTALARNEWEKGNYEKATW
YFGQAMHYDPNNAEAKQDLGNAKQKQG

>**PLC-3 gene**
GGAATTCCATATGGGGAACTCCGTCCGGAAGTTCACTGCATTAGCACGGAACGAATGGGA
AAAGGGGAACTACGAAAAGGCAACTTGGTACTTCGGGCAGGCAATGCACTACGACCCTAA
CAACGTTCGTAAGTTCACGGCCTTGGCCCGTAATGAGTGGGAGAAGGGTAATTACGAGAA
GGCCACGTGGTACTTCGGTCAAGCCATGCACTACGACCCAAATAATGTGCGCAAATTTAC
CGCGCTGGCGCGCAACGAATGGGAAAAAGGCAACTATGAAAAAGCGACCTGGTATTTTGG
CCAGGCGATGCATTATGATCCGAACAACGCGGAAGCGAAACAGGACTTGGGTAATGCCAA
GCAAAAGCAAGGTTGACTCGAGCGG

>**PLC-3 protein**
MGNSVRKFTALARNEWEKGNYEKATWYFGQAMHYDPNNVRKFTALARNEWEKGNYEKATW
YFGQAMHYDPNNVRKFTALARNEWEKGNYEKATWYFGQAMHYDPNNAEAKQDLGNAKQKQ
G

Figure 3.13: **The DNA and protein sequences of PLC derived α-hairpin.** The colour codes indicates N cap in cyan, α-hairpin in green, turn in magenta and solvating helix sequences in yellow. For DNA sequences the NdeI and XhoI restriction sites are displayed with underline. Additional nucleotides (black) required by the restriction enzymes for effective digestion are also displayed at both the termini of the genes

more than one copy of the repeat are produced by assigning the second or third most favoured codon usage for the α-hairpin amino acid sequence. The DNA and protein sequences are shown in Figure 3.14

3.1.6.5 Design of BGA-1, BGA-2 and BGA-3

Glucoamylase (1,4-α-D-glucan glucohydrolase, EC 3.2.1.3,) releases β-D-glucose from the non-reducing ends of starch, glycogen, and malto-oligosaccharides, cleaving all α-glycosidic bonds between glucosyl residues except that of α,α-trehalose. Glucoamylases occur in some prokaryotic and many eukaryotic microorganisms, and may have originated as a polysaccharide exo-hydrolase early in the evolution of glycogen metabolism. Bacterial glucoamylase (BGA) from the thermophilic clostridial species *Thermoanaerobacterium thermosaccharolyticum* consists of two distinct domains, the N-terminal β-domain containing residues 1-251 connected by a short linker domain containing residues 251-295 to a C-terminal α-domain containing residues 296-684 [159]. An α-hairpin containing residues 452-484 in the C-terminal α-domain of HSC resembles the repeating unit of TPR.

Table 3.10: **Homologs of bacterial glucoamylase (BGA). The α-hairpins from the set of 17 BGA proteins are listed in the order from best to worst based on their single repeat P-values for being TPR repeating unit. The idealized TPR unit[§] sequence is given for easy comparison.**

GI number	Sequence
Idealized TPR[§]	`AEAWYNLGNAYYKQGDYDEAIEYYQKALELDPNN`
8777462	`ILTTLWLAQLYIKQGRIKKALNHLKWVVDHRTDL`
16081473	`IITTLWMARYYMRFGDFEKAWNLIQWVKSHRQKS`
15921050	`IITTLWLAEYYLDLGQREKALDYINWAMSRALPS`
23020506	`VLATLWVALYYIEIKEYEKAKDYLRWATKSCTAL`
28948480	`SRDLYHVANAFIAAGDVDSANRSLDYLAKVVKDN`
28948477[¶†]	`SRDLYHVANAFIAAGDVDSANRSLDYLAKVVKDN`
3243238	`SRDLYHVANAFIAAGDVDSANRSLDYLAKVVKDN`
231542	`SRDLYHVANAFIAAGDVDSANRSLDYLAKVVKDN`

Continued on next page

Table 3.10 – continued from previous page

GI number	Sequence
24374006	`PRDFYQCAMAFLAMGDTQTPKVAFEYLKKVQVSD`
15899215	`FISTLWLSQVYSLMGEKDKAKEKIDWVLSKSLPT`
22405949	`IITTLWMARYYIRIGNLEKSWSLIEWVKSHRQKS`
34146785	`SRDLYHVANAFIVAGDTDSANRALDYLDKVVKDN`
17229078	`PRDFYQAAMALLALGDKETPLTAFKYLPQVQVQS`
15922338	`FITTLWLAQQYILEGNKEKAKKYIDWVISHMLPT`
13541259	`IITTLWMARYYLRYGDFERAWDLIMWVKSHRQKS`
14324643	`IITTLWMARYYLRYGDFERAWDLIMWVKSHRQKS`
20808230	`ARDLYHIANAFIAAKDIDSANRALDFLAMVVEKN`

§ Consensus sequence for idealized TPR from Regan *et al.*
† Finally selected α-hairpin
¶ Query sequence hairpin

We extracted a set of 17 homologs of HSC conservatively with BLAST (E-value cutoff of 10^{-3}) against the non-redundant database, using the full length sequence of BGA [GI:28948477] as a query sequence. Each of these sequences were feeded to TPRpred program to find out their sequence fitness to the known TPR units. Results from these comparisons are tabulated in Table 3.10, which shows that BGA from *Thermoactinomyces vulgaris* was ranked best. Family of BGA sequences fall into four major subfamilies that is reflected in the sequences of the α-hairpins, therefore we have selected the query α-hairpin as the best because of its known structure.

Finally, BGA α-hairpin amino acid sequences with necessary features were converted to a DNA sequences according to the most favoured codon usage in *Escherichia coli* using a Perl script. Different nucleotide sequence that codes for more than one copy of the repeat are produced by assigning the second or third most favoured codon usage for the α-hairpin amino acid sequence. The DNA and protein sequences are shown in Figure 3.15.

3.1.7 Model structure of 3 repeat proteins

Comparative models are useful to get a rough idea where the alpha carbons of key residues sit in the folded protein. They can guide mutagenesis experiments,

>**HSC-1 gene**
GGAATTCCATATGGGCAACAGCATTAAAACCCGCAGCCAGCTGATGGTGCAGCAGCTGGA
TGAACAGCAGTGGGAACAGGCGGCGGATACCGTGCGCAAACTGCGCTTTCTGGACCCGAA
CAACGCGGAAGCGAAACAGGACTTGGGTAATGCCAAGCAAAAGCAAGGTTGACTCGAGCG
G

>**HSC-1 protein**
MGNSIKTRSQLMVQQLDEQQWEQAADTVRKLRFLDPNNAEAKQDLGNAKQKQG

>**HSC-2 gene**
GGAATTCCATATGGGTAATTCGATCAAGACGCGTTCGCAATTGATGGTTCAACAATTGGA
CGAGCAACAATGGGAGCAAGCCGCCGACACGGTTCGTAAGTTGCGTTTCTTGGACCCAAA
TAATATTAAAACCCGCAGCCAGCTGATGGTGCAGCAGCTGGATGAACAGCAGTGGGAACA
GGCGGCGGATACCGTGCGCAAACTGCGCTTTCTGGACCCGAACAACGCGGAAGCGAAACA
GGACTTGGGTAATGCCAAGCAAAAGCAAGGTTGACTCGAGCGG

>**HSC-2 protein**
MGNSIKTRSQLMVQQLDEQQWEQAADTVRKLRFLDPNNIKTRSQLMVQQLDEQQWEQAAD
TVRKLRFLDPNNAEAKQDLGNAKQKQG

>**HSC-3 gene**
GGAATTCCATATGGGGAACTCCATTAAGACTCGGTCCCAGTTAATGGTCCAGCAGTTAGA
CGAACAGCAGTGGGAACAGGCAGCAGACACTGTCCGGAAGTTACGGTTCTTAGACCCTAA
CAACATCAAGACGCGTTCGCAATTGATGGTTCAACAATTGGACGAGCAACAATGGGAGCA
AGCCGCCGACACGGTTCGTAAGTTGCGTTTCTTGGACCCAAATAATATTAAAACCCGCAG
CCAGCTGATGGTGCAGCAGCTGGATGAACAGCAGTGGGAACAGGCGGCGGATACCGTGCG
CAAACTGCGCTTTCTGGACCCGAACAACGCGGAAGCGAAACAGGACTTGGGTAATGCCAA
GCAAAAGCAAGGTTGACTCGAGCGG

>**HSC-3 protein**
MGNSIKTRSQLMVQQLDEQQWEQAADTVRKLRFLDPNNIKTRSQLMVQQLDEQQWEQAAD
TVRKLRFLDPNNIKTRSQLMVQQLDEQQWEQAADTVRKLRFLDPNNAEAKQDLGNAKQKQ
G

Figure 3.14: **The DNA and protein sequences of HSC derived α-hairpin.** The colour codes indicates N cap in cyan, α-hairpin in green, turn in magenta and solvating helix sequences in yellow. For DNA sequences the NdeI and XhoI restriction sites are displayed with an underline. Additional nucleotides (black) required by the restriction enzymes for effective digestion are also displayed at both the termini of the genes

>**BGA-1 gene**

GGAATTCCATATGGGCAACAGCAGCCGCGATCTGTATCATGTGGCGAACGCGTTTATTGCGGCGGGCGATGTGGATAGCGCGAACCGCAGCCTGGATTATCTGGCGAAAGTGGACCCGAACAACGCGGAAGCGAAACAGGACTTGGGTAATGCCAAGCAAAAGCAAGGTTGACTCGAGCGG

>**BGA-1 protein**

MGNSSRDLYHVANAFIAAGDVDSANRSLDYLAKVDPNNAEAKQDLGNAKQKQG

>**BGA-2 gene**

GGAATTCCATATGGGTAATTCGTCGCGTGACTTGTACCACGTTGCCAATGCCTTCATCGCCGCCGGTGACGTTGACTCGGCCAATCGTTCGTTGGACTACTTGGCCAAGGTTGACCCAAATAATAGCCGCGATCTGTATCATGTGGCGAACGCGTTTATTGCGGCGGGCGATGTGGATAGCGCGAACCGCAGCCTGGATTATCTGGCGAAAGTGGACCCGAACAACGCGGAAGCGAAACAGGACTTGGGTAATGCCAAGCAAAAGCAAGGTTGACTCGAGCGG

>**BGA-2 protein**

MGNSSRDLYHVANAFIAAGDVDSANRSLDYLAKVDPNNSRDLYHVANAFIAAGDVDSANRSLDYLAKVDPNNAEAKQDLGNAKQKQG

>**BGA-3 gene**

GGAATTCCATATGGGGAACTCCTCCCGGGACTTATACCACGTCGCAAACGCATTCATTGCAGCAGGGGACGTCGACTCCGCAAACCGGTCCTTAGACTACTTAGCAAAGGTCGACCCTAACAACTCGCGTGACTTGTACCACGTTGCCAATGCCTTCATCGCCGCCGGTGACGTTGACTCGGCCAATCGTTCGTTGGACTACTTGGCCAAGGTTGACCCAAATAATAGCCGCGATCTGTATCATGTGGCGAACGCGTTTATTGCGGCGGGCGATGTGGATAGCGCGAACCGCAGCCTGGATTATCTGGCGAAAGTGGACCCGAACAACGCGGAAGCGAAACAGGACTTGGGTAATGCCAAGCAAAAGCAAGGTTGACTCGAGCGG

>**BGA-3 protein**

MGNSSRDLYHVANAFIAAGDVDSANRSLDYLAKVDPNNSRDLYHVANAFIAAGDVDSANRSLDYLAKVDPNNSRDLYHVANAFIAAGDVDSANRSLDYLAKVDPNNAEAKQDLGNAKQKQG

Figure 3.15: **The DNA and protein sequences of BGA derived α-hairpin.** The colour codes indicates N cap in cyan, α-hairpin in green, turn in magenta and solvating helix sequences in yellow. For DNA sequences the NdeI and XhoI restriction sites are displayed with an underline. Additional nucleotides (black) required by the restriction enzymes for effective digestion are also displayed at both the termini of the genes

or hypotheses about structure-function relationships. Therefore, we modeled the 3 repeat proteins from each of the five hairpins selected, and are shown in Figure 3.16 and Figure 3.17.

The angle between the two helices of a single repeat unit, angle between the two juxtaposed repeat units, curvature of the 3-repeat superhelix, radius of the superhelical axis, and the direction of the polypeptide chain along the superhelical axis are comparable with those of known and designed TPR structures. From this it is obvious that these proteins have the propensity to form the stacked TPR units.

3.2 Experimental analysis

To evaluate the expression ability, extent of solubility, foldability and stability of internally duplicated α-hairpins, we constructed the semi-synthetic genes which place each of the 15 genes under the control of the T7 promoter, with a N-terminal 6× his tag. The predicted biophysical properties of these 15 proteins are tabulated in Table 3.11.

There are 20^{100} possible sequences for a domain of 100 residues, and only a negligible fraction of these will be able to fold, let alone display a biological activity. Therefore, we tested the expression ability and solubility of each of these 15 proteins. Results of these experiments are shown in Figures 3.18A-3.22A. Following induction in *E. coli*, each of the proteins are clearly visible on Coomassie-stained tricine-urea-SDS-PAGE gels as shown in Figure 3.18A-3.22A. Further, we tested their extent of solubility by analysing the soluble and insoluble fraction of the cell lysate on Coomassie-stained tricine-urea-SDS-PAGE gels, as shown in Figures 3.18B-3.22B. The proteins derived from the Tom20 and RPS20 α-hairpins were soluble, whereas the proteins derived from the PLC, HSC and BGA α-hairpins were not soluble. Although BGA-3 proteins was 50 % soluble, this fraction was aggregating during the later steps of purification.

Soluble Tom20-1, Tom20-2, Tom20-3, RPS20-1, RPS20-2 and RPS20-3 proteins survive proteolysis in the cell. However, prolonged storage of cell lysate and also purified protein samples, without protease inhibitors, did lead to degradation to some extent. From a protein design point of view, the fact that the these proteins were sufficiently stable to survive degradation by endogenous pro-

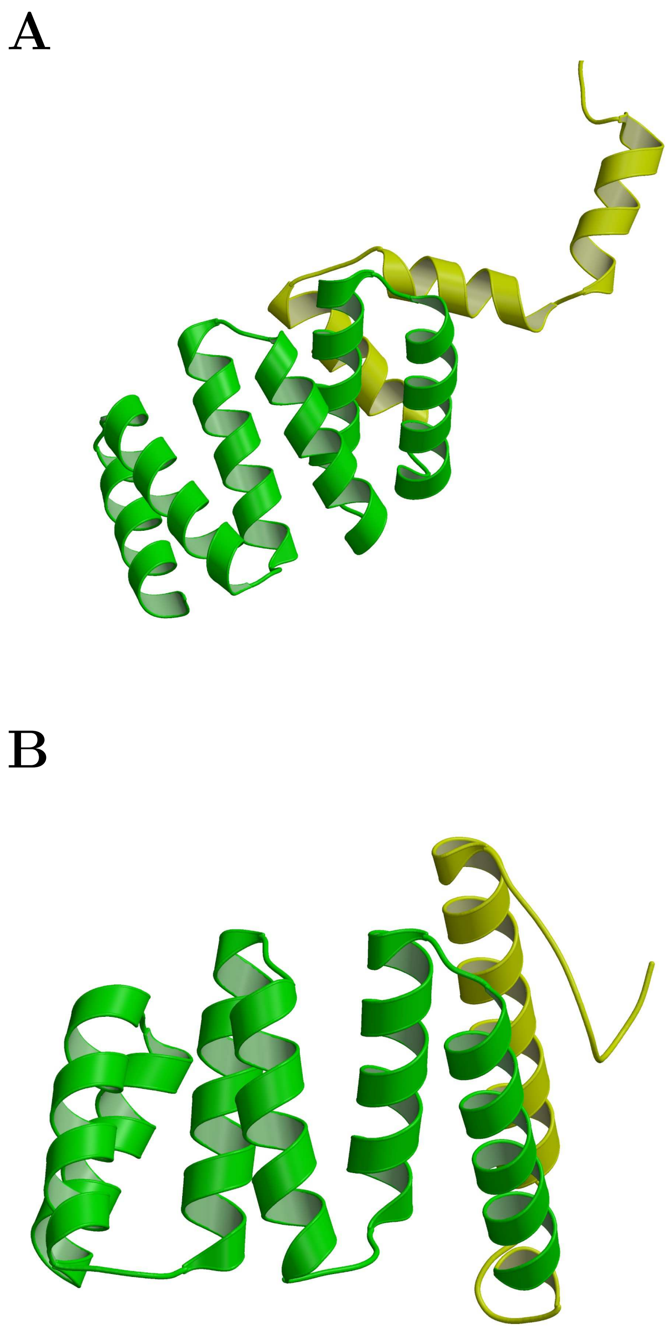


Figure 3.16: **Model of 3 repeat proteins with natural solvating helix.** Triplicated TPR-like α-hairpins are shown in green and the ”solvating” helices are shown in yellow. (A)Tom20-3. (B) RPS20-3

.

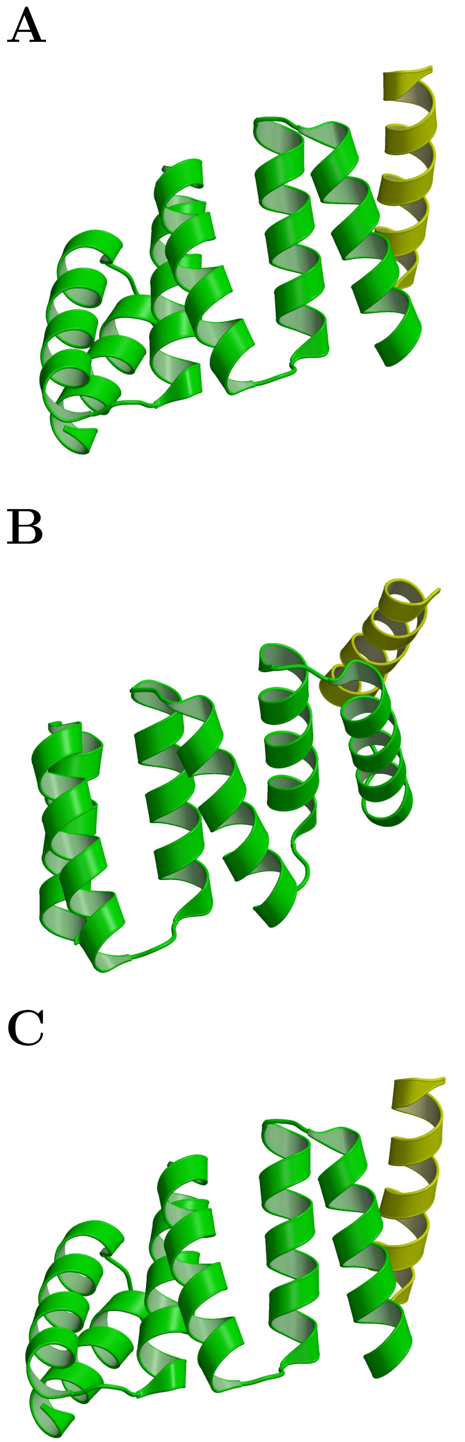


Figure 3.17: **Model of 3 repeat proteins without natural solvating helix.** Triplicated TPR-like α-hairpins are shown in green and the "solvating" helices are shown in yellow. (A) PLC-3. (B) HSC-3. (C) BGA-3.

teases [160,161] is a key outcome given that even some natural proteins are easily degraded when recombinantly expressed in *E. coli* [162, 163]. Degradation by endogenous proteases is a defence mechanism used by cells against proteins that misfold and/or are particularly unstable.

These proteins could also be readily purified using the combination of ion-exchange, nickel-affinity and gel-sizing chromatography and stay folded sufficiently not to form aggregates during purification steps. This is demonstrated by gel filtration experiments, which provided no evidence of these proteins present in the void volume. In particular, a single sharp peak was observed for each of these proteins and were also corresponding to a monomeric proteins as estimated from a calibration plot (data not shown). It was also possible to purify the insoluble proteins to reasonable homogeneity by nickel-affinity chromatography. However, they were proven to aggregate during refolding. The folding and stability of several of these artificial proteins were assessed using various biophysical techniques.

Table 3.11: **Predicted biophysical properties of the proteins.** All the proteins contains additional 20 amino acids (2.1 KDa) as a consequence of the cloning strategy that includes six histidine residues that are used as a tag.

Parent	**Name**	**Mol. Wt.**	**PI**
Tom20	Tom20-1	11	5.47
	Tom20-2	14.7	4.77
	Tom20-3	18.4	4.50
RPS20	RPS20-1	10.4	11.11
	RPS20-2	14.2	10.48
	RPS20-3	17.9	10.26
PLC	PLC-1	8.3	9.46
	PLC-2	12.4	9.40
	PLC-3	16.5	9.37
HSC	HSC-1	8.2	9.40
	HSC-2	12.3	9.22
	HSC-3	16.4	9.10
BGA	BGA-1	7.8	7.10

Continued on next page

Table 3.11 – continued from previous page

Parent	Name	Mol. Wt. KDa	pI
	BGA-2	11.5	6.48
	BGA-3	15.2	6.16

3.2.1 Biochemical characterization

3.2.1.1 CD spectroscopy

The CD spectra of Tom20-1, Tom20-2 and Tom20-3 proteins indicate significant α-helical content, with minima at 222 nm and 207 nm, as illustrated in Figure 3.23A. However, the minima shows a modest shift upon insertion of two and three repeats. Although the modest shift could be due to destabilization, the overall similarity of the far-UV CD spectra of Tom20-2 and Tom20-3 proteins is in consistence with a typical α-helical protein. Similar behaviour was also observed for the RPS20-1, RPS20-2 and RPS20-3 proteins as illustrated in Figure 3.23B. These proteins appear to have native-like properties as assessed by other biophysical experiments.

The double minima for PLC-2 and PLC-3 proteins were not pronounced, PLC-1 appears to have a minimum at 207 nm as illustrated in Figure 3.24A. The basis for this anomalous behaviour is unclear. The HSC-1, HSC-2 and HSC-3 proteins appear to have modest α-helical content and the typical α-helical behaviour could be recovered for HSC-3 by adding 10 % TFE. This shows that HSC-1, HSC-2 and HSC-3 are more like a molten-globule (Figure 3.24B). The CD spectra for the BGA-3 shows a clear random-coil behaviour. However, the α-helical behaviour could be recovered with 20 % TFE (Figure 3.24C). Moreover, BGA-1, BGA-2 and BGA-3 proteins visibly aggregates at low concentrations, therefore BGA-1, BGA-2 and BGA-3 proteins could not be further characterized.

3.2.1.2 Thermal denaturation

The common features of many naturally occurring proteins is co-operative thermal denaturation transitions. This feature appears to be linked to the rigidity of the protein structure. A highly cooperative thermal transition indicates a large

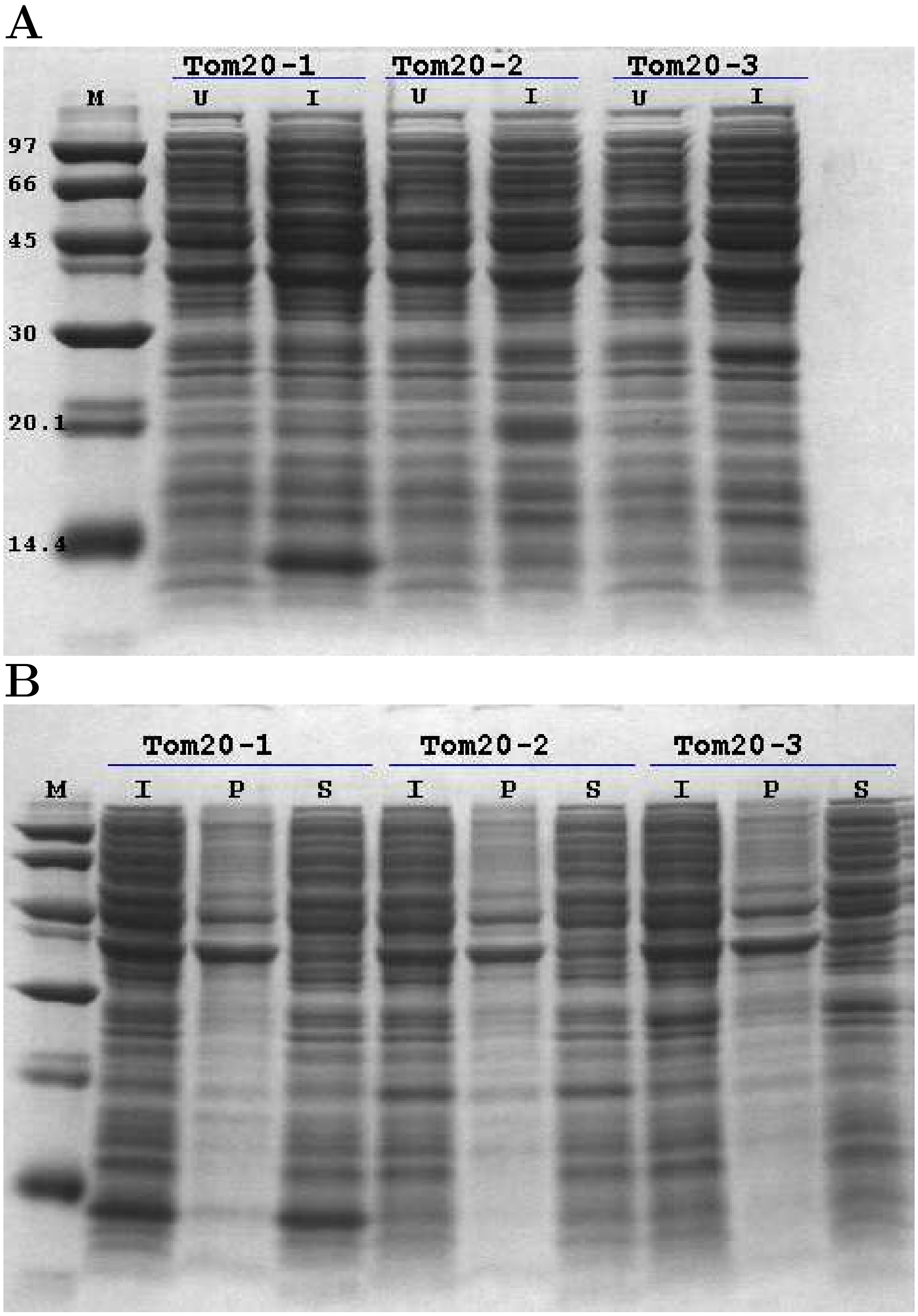


Figure 3.18: **Test expression and solubility of Tom20-1, Tom20-2 and Tom20-3 proteins**. The names of the proteins are shown on the top of the gel, the two or three lines for each protein represent the uninduced (U), induced (I), insoluble fraction (P) and the soluble fraction (S) of the cell lysate. (A) Test expression. (B) Solubility test.

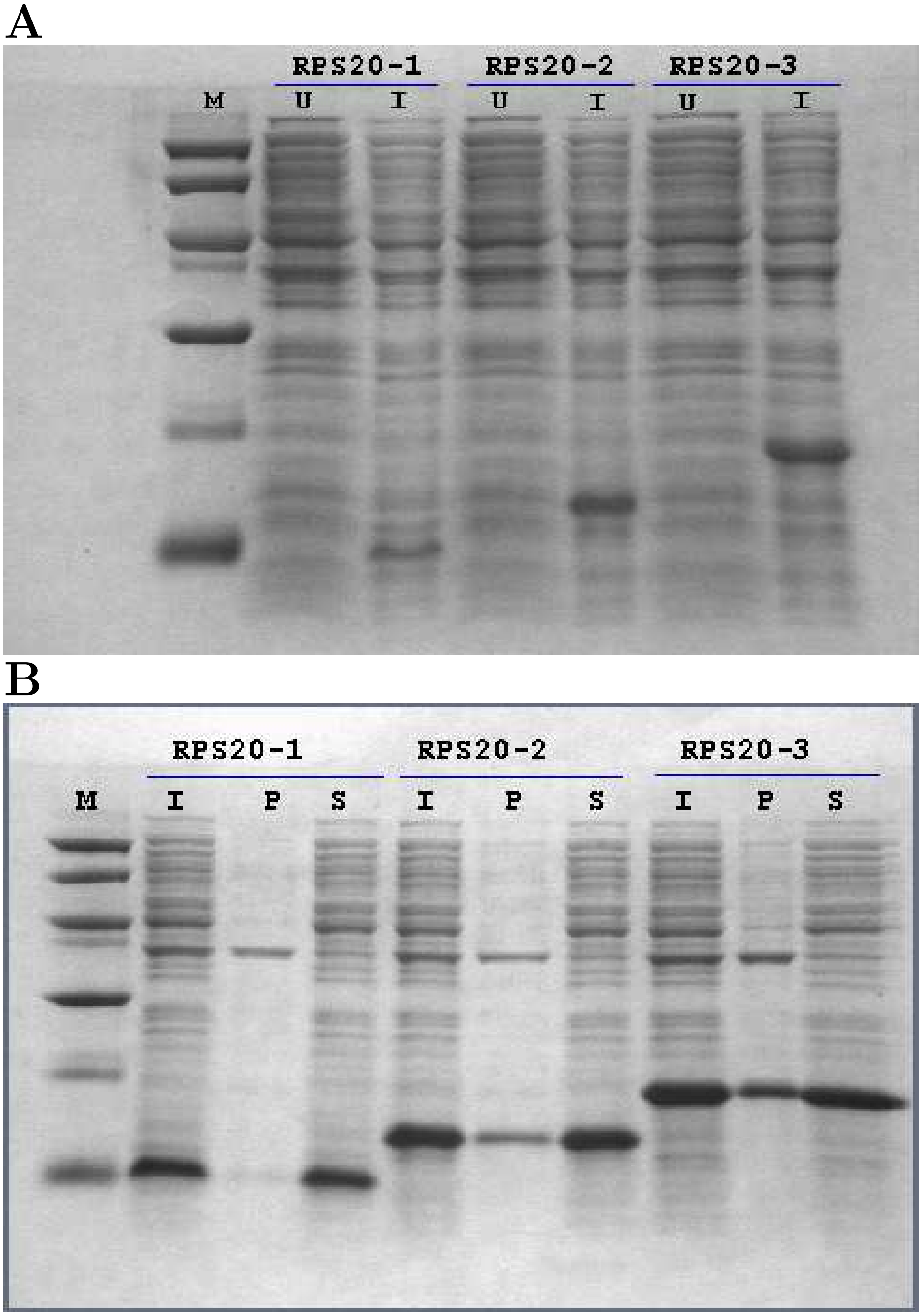


Figure 3.19: **Test expression and solubility of RPS20-1, RPS20-2 and RPS20-3 proteins**. The names of the proteins are shown on the top of the gel, the two or three lines for each protein represent the uninduced (U), induced (I), insoluble fraction (P) and the soluble fraction (S) of the cell lysate. (A) Test expression. (B) Solubility test.

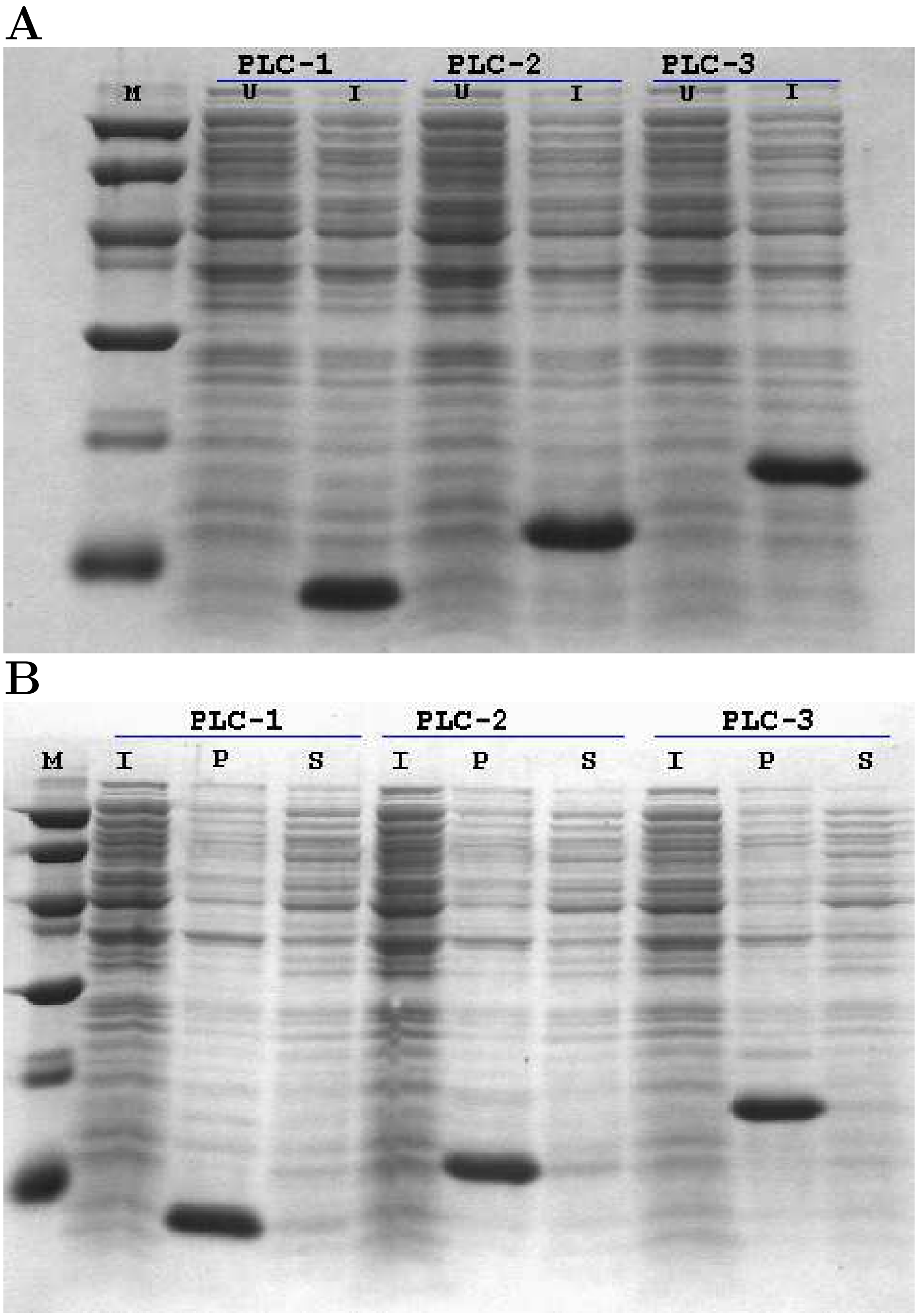


Figure 3.20: **Test expression and solubility of PLC-1, PLC-2 and PLC-3 proteins**. The names of the proteins are shown on the top of the gel, the two or three lines for each protein represent the uninduced (U), induced (I), insoluble fraction (P) and the soluble fraction (S) of the cell lysate. (A) Test expression. (B) Solubility test.

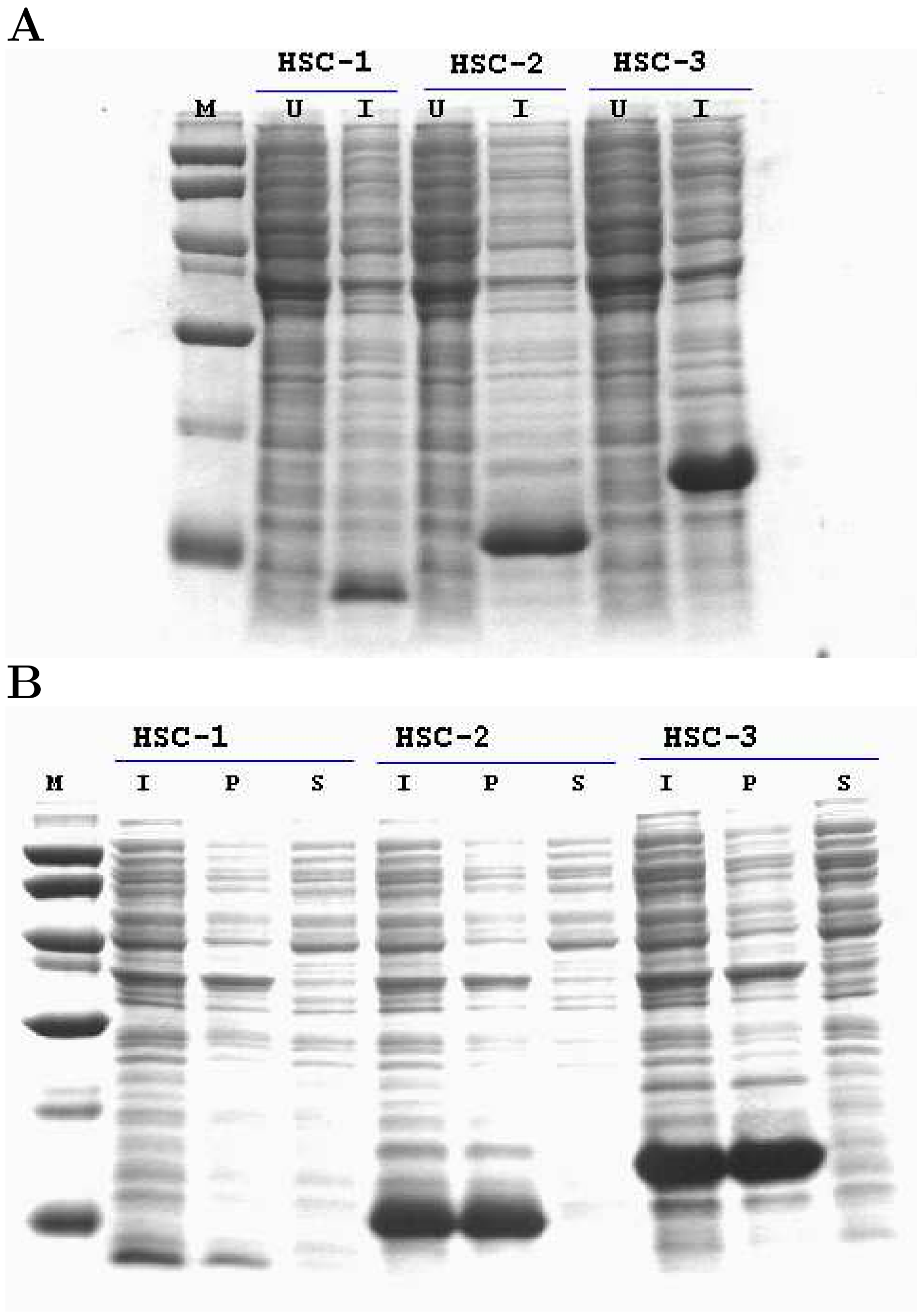


Figure 3.21: **Test expression and solubility of HSC-1, HSC-2 and HSC-3 proteins**. The names of the proteins are shown on the top of the gel, the two or three lines for each protein represent the uninduced (U), induced (I), insoluble fraction (P) and the soluble fraction (S) of the cell lysate. (A) Test expression. (B) Solubility test.

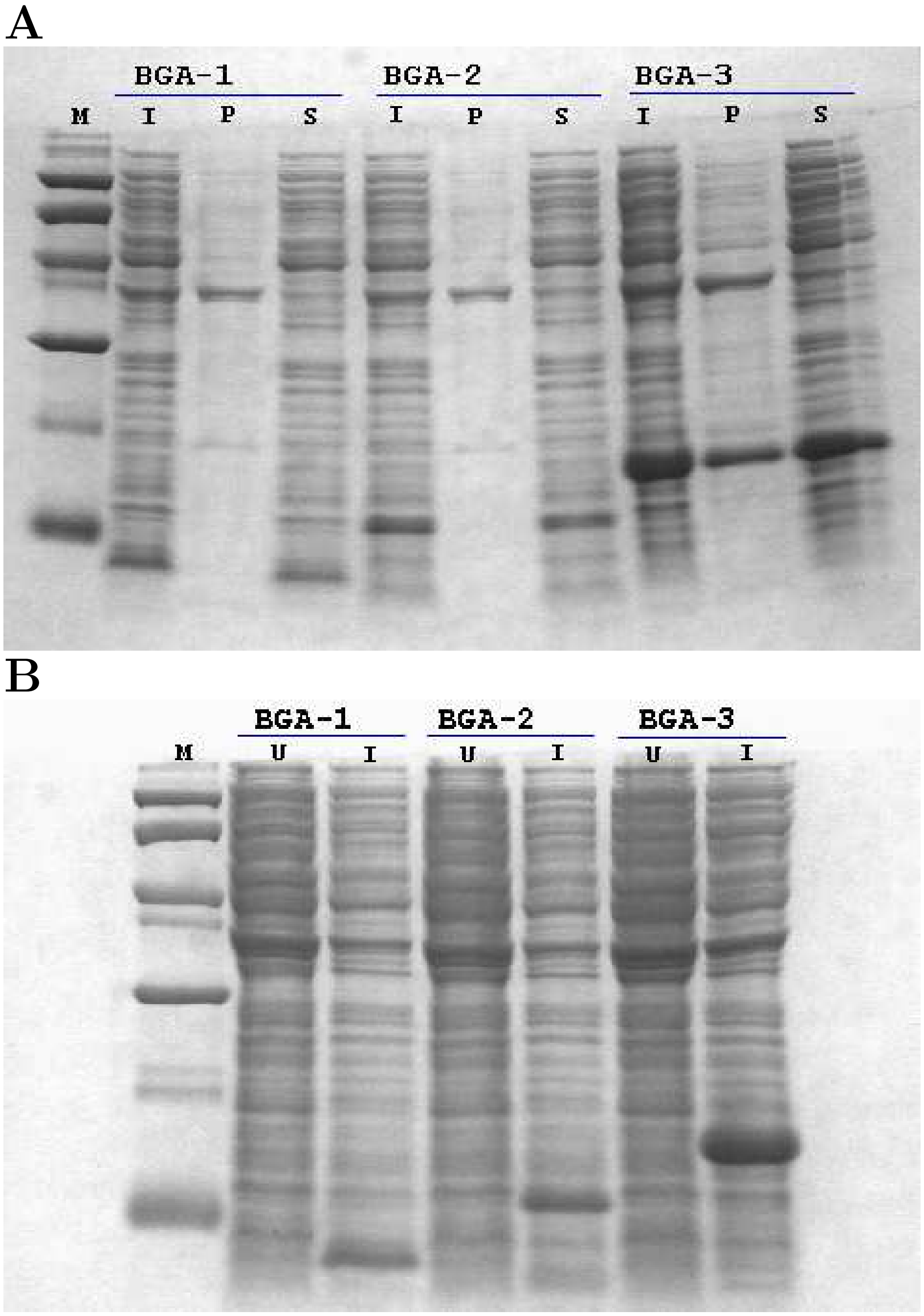


Figure 3.22: **Test expression and solubility of BGA-1, BGA-2 and BGA-3 proteins**. The names of the proteins are shown on the top of the gel, the two or three lines for each protein represent the uninduced (U), induced (I), insoluble fraction (P) and the soluble fraction (S) of the cell lysate. (A) Test expression. (B) Solubility test.

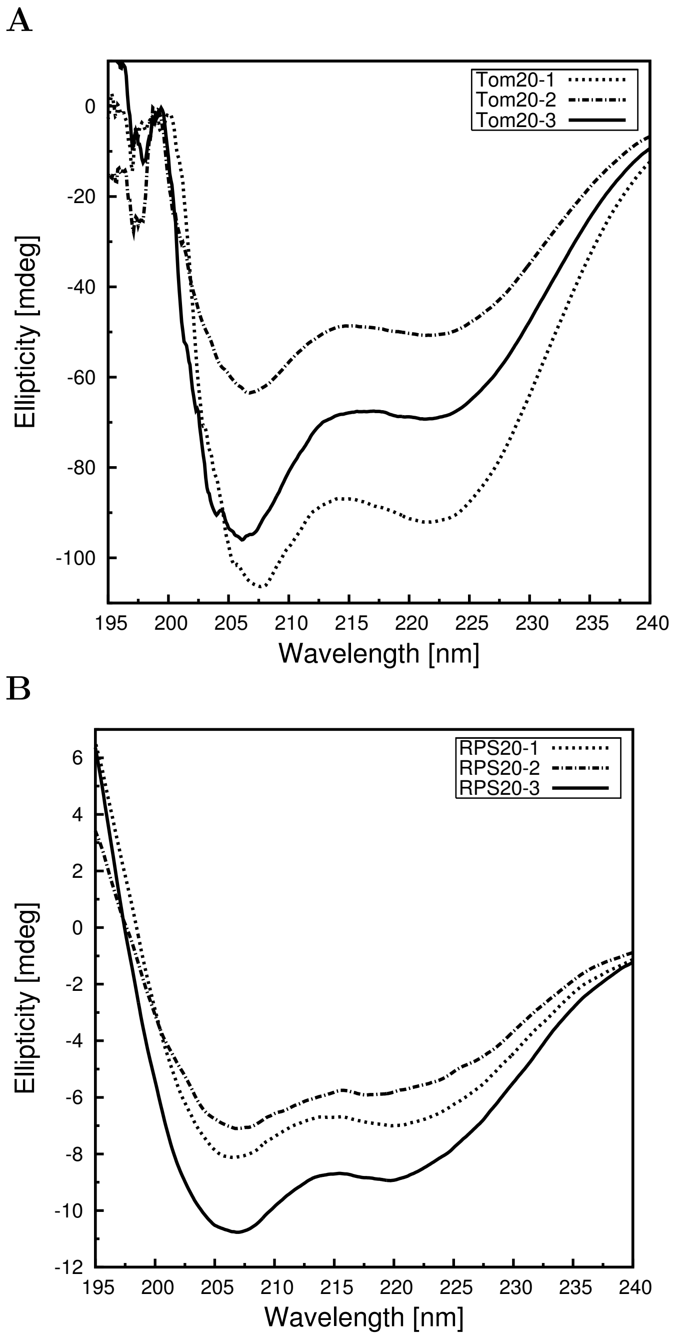


Figure 3.23: **Circular dichroism spectra of the derived artificial proteins.** The CD spectra of the artificial proteins shows the expected secondary structure content. (A) Tom20-1, Tom20-2 and Tom20-3. (B) RPS20-1, RPS20-2 and RPS20-3.

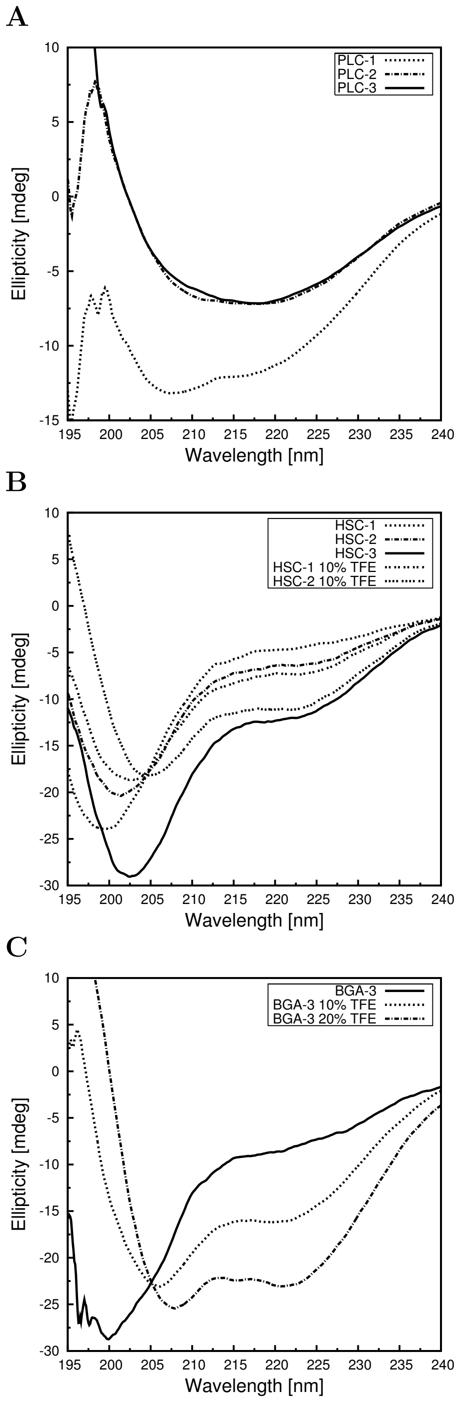


Figure 3.24: **Circular dichroism spectra of the derived artificial proteins.** The CD spectra of the artificial proteins shows anomalous and random-coil behaviour. (A) PLC-1, PLC-2 and PLC-3 proteins. (B) HSC-1, HSC-2 and HSC-3 proteins with and without TFE. (C) BGA-3 protein with and without TFE.

change in enthalpy upon unfolding and is consistent with a change from a rigid folded protein to a dynamic unfolded protein.

To examine the thermodynamic stability of each of these proteins, temperature-induced unfolding transitions were measured by monitoring the change in the ellipticity at 222 nm as a function of temperature (Figure 3.25). Tom20-1, Tom20-2 and Tom20-3 proteins exhibit cooperative unfolding transitions, with midpoints of thermal denaturations of 77 °C (Tom20-1), 67 °C (Tom20-2), and 77 °C (Tom20-3) and reasonable reversibility (data not shown). RPS20-1, RPS20-2 and RPS20-3 proteins exhibit co-operative unfolding transitions, with midpoints of the thermal denaturations of 45 °C (RPS20-1), 50 °C (RPS20-2), and 53 °C (RPS20-3) and reasonable reversibility (data not shown). We could not demonstrate co-operative unfolding transitions for other proteins.

3.2.1.3 Trp fluorescence

In order to determine the presence of tertiary structure in PLC-1, PLC-2, PLC-3, HSC-1, HSC-2 and HSC-3 proteins, as they contains Trp residues, we used Trp fluorescence spectroscopy. The Trp maximum emission wavelength was at 349 nm for the native PLC-2 and PLC-3, which indicates a partial exposure of the Trp indole ring to an aqueous environment (Figure 3.26). Upon denaturing by adding 8 M urea, the maximum wavelength of the Trp emission blue shifted to around 355 nm, which suggests that the Trp side chain is now in a much more solvent accessible hydrophilic environment. When we compare the fluorescence spectrum of the native PLC-2 and PLC-3 and denatured PLC-2 and PLC-3, we observe an increase of the fluorescence in the PLC-2/3 by ∼2 fold. This suggests that the Trp fluorescence in the native proteins is quenched.

The Trp maximum emission wavelength was at 350 nm for the native HSC-3, which indicates a full exposure of the Trp indole ring to an aqueous environment. However, the Trp maximum emission wavelength was shifted to 340 nm in presence of 10 % TFE, which suggests that the protein begins to fold. This was further evidenced by decrease of the fluorescence, as illustrated in Figure 3.26.

The Trp fluorescence of PLC and HSC derived proteins is not strong enough to show these proteins are folded.

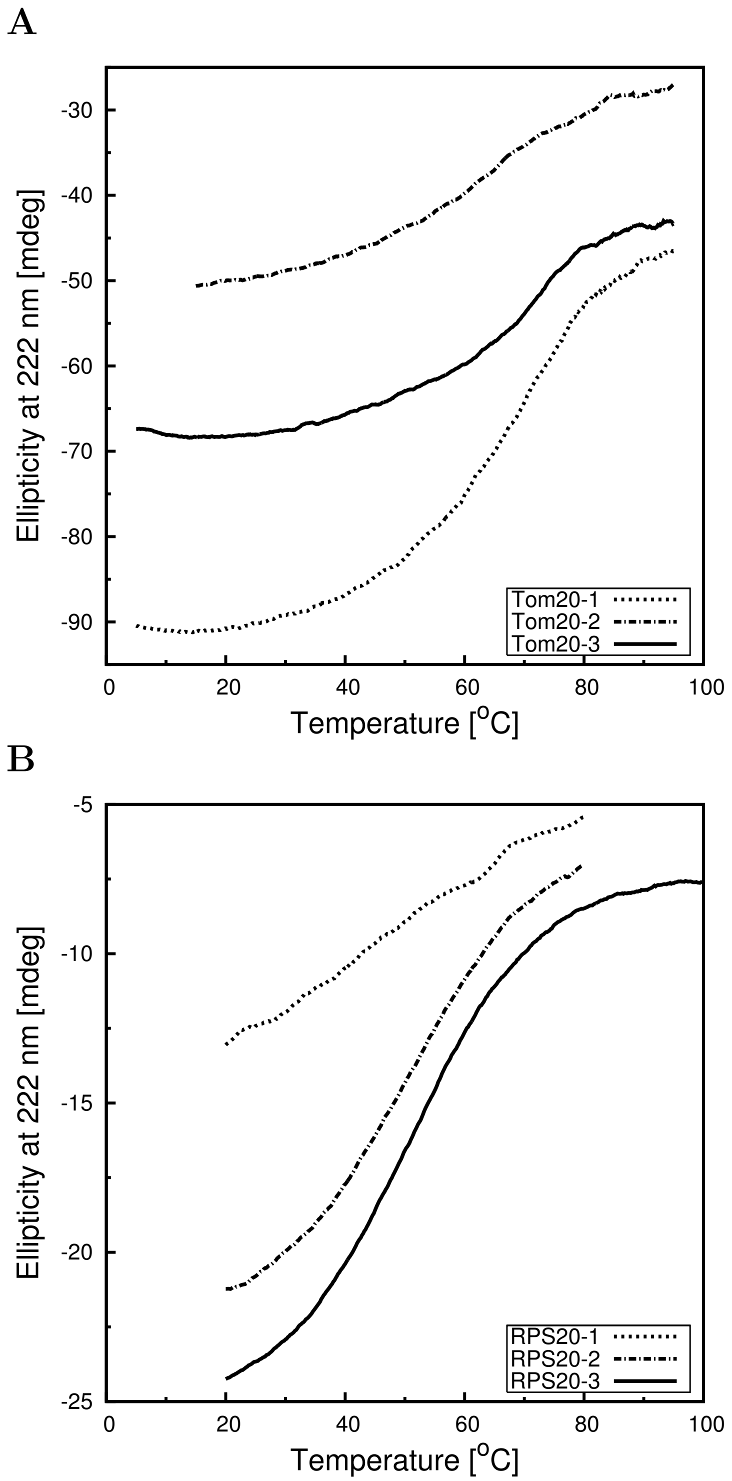


Figure 3.25: **Thermal denaturation of the derived artificial proteins.** The temperature-induced denaturation profiles exists in two-state and are co-operative. (A) Tom20-1, Tom20-2 and Tom20-3 proteins. (B) RPS20-1, RPS20-2 and RPS20-3 proteins.

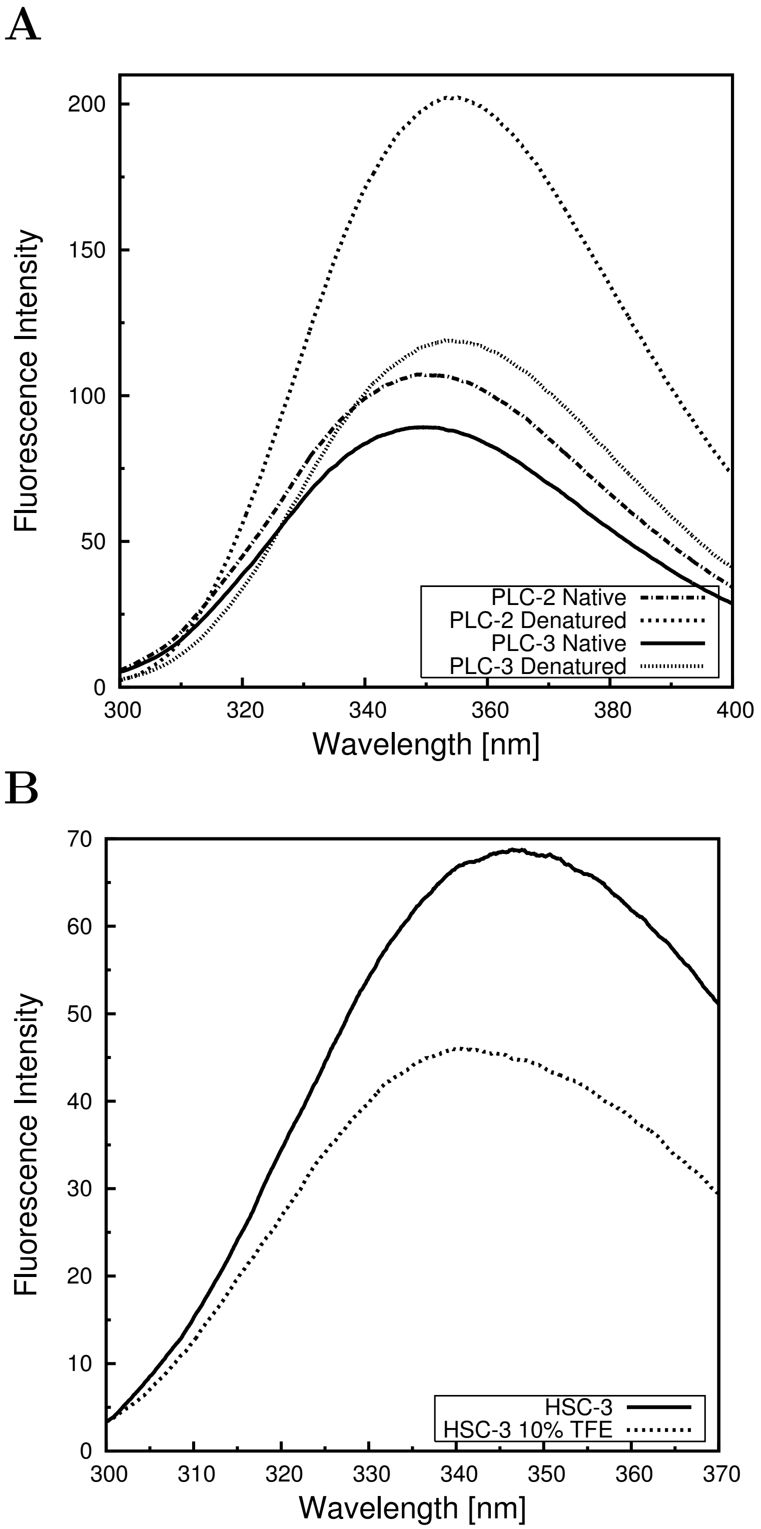


Figure 3.26: **Trp fluorescence spectroscopy of the derived artificial proteins.** (A) PLC-1, PLC-2 and PLC-3. (B) HSC-3 protein.

3.2.2 In vitro chaperone activity of Tom20-1/2/3

The Tom20 protein acts as a mitochondrial import receptor which recognizes presequence and facilitates protein import into mitochondria. The chaperone-like activity was reported previously for the cytosolic domain of natural Tom20 protein [164]. However, it is reasonable to assume that, like Tom20, Tom20-2 and Tom20-3 proteins are involved in binding to unfolded proteins. We, therefore tested Tom20-1 Tom20-2 and Tom20-3 for their ability to interact with unfolded proteins. The unfolded porcine citrate synthase during heat-induced denaturation aggregates. Resulting aggregation of this protein was measured as increase in attenuance caused by light scattering. When Tom20-1, Tom20-2 and Tom20-3 were present in stoichiometric concentrations during the heat-incubation period, aggregation of citrate synthase was suppressed, demonstrating a basic chaperone activity of Tom20-1, Tom20-2 and Tom20-3.

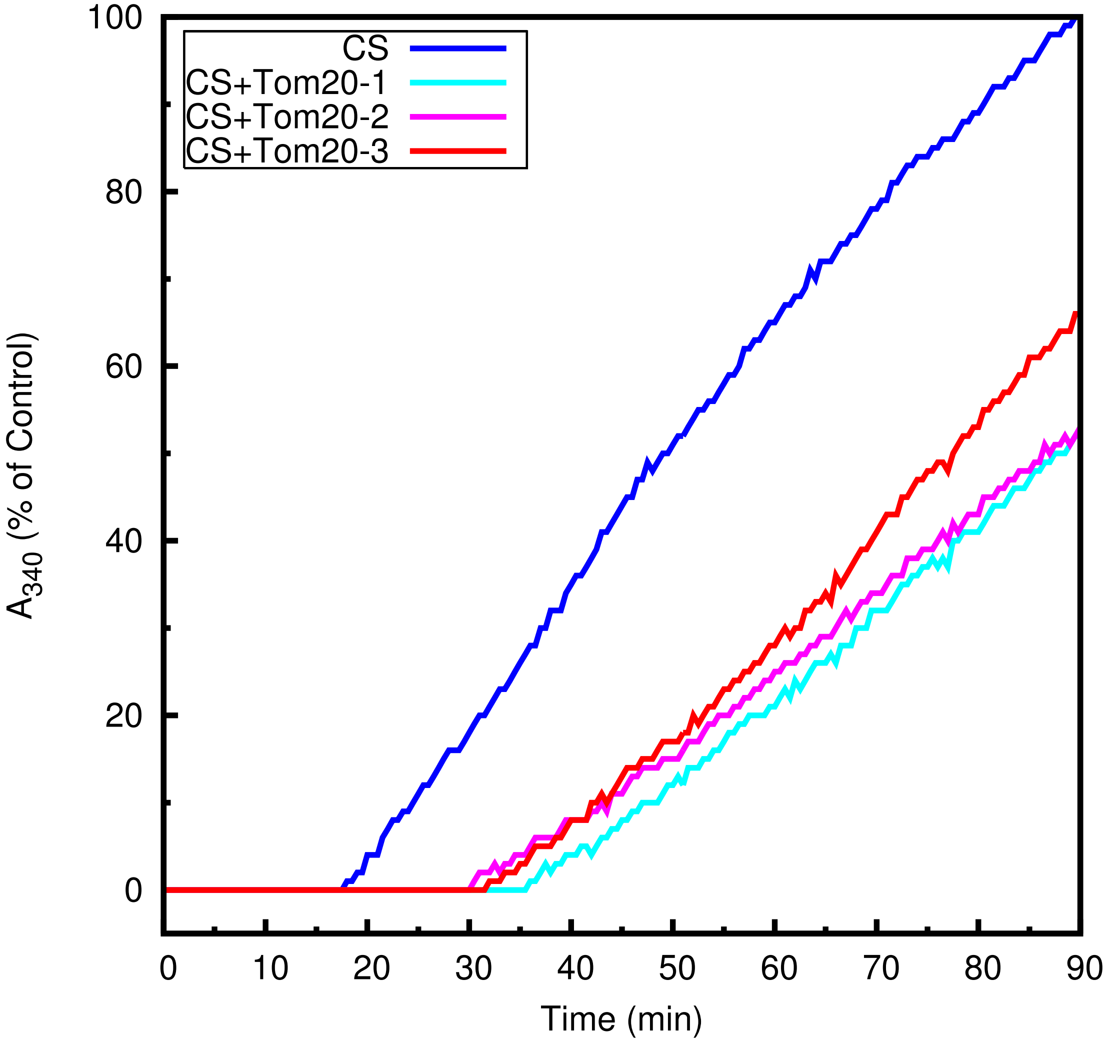


Figure 3.27: **Chaperone activity of Tom20-1, Tom20-2 and Tom20-3 proteins**. Heat-inducible aggregation of citrate synthase(CS) at 45 °C in the absence or presence of a fivefold molar excess of Tom20-1, Tom20-2 and Tom20-3.

Chapter 4

Discussion

4.1 The detection of TPRs

The detection of homologous repeating units of TPR from solenoid repeat proteins is challenging because of less similarity at sequence level. Currently available resources like Pfam, SMART and REP uses their own database of profile hidden Markov models (HMMs) or sequence profiles which are constructed from known repeat families. However, these profiles are constructed from closely homologous repeats; therefore, divergent repeat units often get a negative score and are not considered in computing the overall statistical significance, even though they are individually significant.

The general database search methods like BLAST and HMMER methods which estimates P-values for repeat proteins are not explicitly using the distributions of scores for suboptimal alignments. On the other hand, REP used for the detection of solonoid repeats estimates the single repeat P-value by using the score distribution of sub-optimal non-overlapping alignment scores of a randomised sequence database but follows Gumbel distribution where the tails approach zero much slower, therefore the P-value computed will be less significant.

We demonstrate that for TPRs by not allowing internal gaps within repeats, the score distribution of the repeat alignment with unrelated sequences follows a near Gaussian distribution thereby, the same positive score of a putative repeat unit will generally have a much higher significance when compared to a Gumbel distribution. Hence, the restriction of ungapped repeats increases the sensitivity for detecting ungapped repeat families such as TPRs, PPRs and SEL1-like repeats. The single repeat P-values can further be used to include divergent repeat units by a P-value-dependent score offset. Putative repeat units located near an already identified repeat get a reward. This approach has been coded into a program called TPRpred.

Generally, distant homologous relationships among the proteins is often established only by comparing the three-dimensional structures of the proteins, because the sequence divergence can be so large that simple comparison of their sequences fails to identify any similarity. New generation homology detection tools use averaged sequences of entire homologous families in profiles to detect such homologies. It is shown that details of profile calculation strongly influence its sensitivity in recognizing distant homologies. The most important choice is how to include information from diverging members of the family, avoiding false predictions, while accounting for entire sequence divergence within a family. PSI-BLAST takes a conservative approach, deriving a profile from core members of the family, providing a solid improvement without almost any false predictions. Therefore we have taken a PSI-BLAST approach in building the optimized profiles through iterative searches by varying the thresholds for inclusion of repeats into the profile. The best profile is selected from the pool of generated profiles based on their performance on proteins of known structure.

The performance of TPRpred is compared with HMMER, a method that employs a different methodology for the detection of repeats. HMMER is the search method used by the Pfam and SMART domain databases. Although the primary goal of HMMER is to identify domains, its capability to detect repeats using emphiric thresholds has drastically improved the annotation quality of repeats in new sequences. Comparison of HMMER with TPRpred demonstrates that, TPRpred detects more sequences with E-value better than the first false positive. However, for lower selectivity TPRpred performance is comparable to HMMER as illustrated in Figure 3.5. The real difference between HMMER based methods, REP and TPRpred became visible on STAND family members, and on the known protein structures as tabulated in Table 3.2 and Table 3.3. For these sets, TPRpred recognized more number of TPR-containing proteins and also detected most of the individual repeats. TPRpred also detected repeats that are missed by other servers as illustrated in Figure 3.6.

The TPR family of repeats occur in extremely diverse phyla and is likely to have originated prior to the last common ancestor of archaea, eukarya and bacteria. TPR sequences are extremely diverse at the sequence level. Although majority of the repeats can be detected using the single sequence profile, Pfam server has divided the TPRs into several families. Our results suggesting that, the unification of existing TPR families of Pfam into a single family.

TPRpred shows a marked improvement over existing methods, particularly in the detection of non-canonical, divergent repeats. We attribute this to the exploitation of simple traits such as the tendency of repeats to occur in tandem, along with robust statistical evaluations and the construction of profiles by iterative searches. The algorithmic improvements of the P-value-dependent score offset as well as the tight-fit reward are quite general and easily transferable to other repeat detection approaches.

4.2 Evolution of protein domains

Protein domains are compact polypeptide structures, generally organized around a clearly recognizable hydrophobic core and associated with a specific function or activity. Eventhough they adopt only to a limited number of folds, it is unclear about the origin of these folds. whether each fold originated just once and propagated via divergent evolution or on multiple occasions by convergent evolution of structures. It is equally unclear whether some seemingly different folds share a common ancestor or whether each arose separately in evolution [165].

A protein fold is a simplified representation of protein structure that was originally intended to be invariant to possible conformational changes. The potential flexibility of protein loops and other peripheral regions, and their structural variability in homologous proteins were recognised a long time ago. Therefore, the fold of a protein was defined by the composition, architecture and topology of its core secondary structural elements (i.e. α helices and/or β strands). The discovery of sequences that can adopt alternative secondary structures in the same protein (chameleon sequences), thereby affecting its composition, have already shaken the presumed invariance of protein fold [166, 167]. Recent structures revealed even more remarkable examples of large-scale fold variations, altering the protein architecture and topology. These examples provide new insights into protein fold evolution and may be of value in the quest for an evolving paradigm of changeable fold [168].

There is an increase in the number of chameleon sequences that adopt two distinct folded conformations in native conditions. Previously, they were thought to represent two different functional states: one is the free protein, while the other is the ligand-bound form. Recently, both conformations were shown to be at equilibrium in ligand-free form [165]. Apart from this, the fold changes due to some genetic events that led to the fundamental changes in the structure of protein domains. Insertions and deletions (indels) together with single amino acid substitutions are the most common events in protein evolution. Indels are about an order of magnitude less frequent than residue substitutions [169, 170]. One might look at indels as largely neutral or deleterious mishaps or as vehicles of progress: the right indel might relax structural tension accumulated in the course of amino acid substitutions. While the true role of indels in protein evolution remains to be investigated, it is quite clear that they offer a way that can potentially lead to significant and even drastic structural changes [171]. In additional to these classical genetic events, new events are discovered recently. Foremost among these events is circular permutation, which presumably occurs by gene duplication, fusion, and partial deletion [172], which can lead to substantial changes in the topology of a protein fold. Evidence for past intragene duplications causing short repetitions within proteins and giving rise to new structural variants is also compelling. Finally, illegitimate recombination, occurring between unrelated genes, also leads to new folds where the recombined parts prove structurally compatible [165].

More fundamental and challenging question is the evolutionary origin of the domain itself. Perhaps the earliest evolutionary unit corresponded to a much smaller structural unit than a domain and the modern domain folds might have evolved from peptide ancestors [165, 79]. By applying structure similarity detection method to detect instances where localized regions of different protein folds contain highly similar sequences and structures, an all-on-all comparison of known structures resulted in the numerous instances of local sequence and structure similarities within different protein folds, together with evidence from proteins containing sequence and structure repeats, favoured the argument of the evolution of modern single polypeptide domains from ancient short peptide ancestors called antecedent domain segments (ADSs). In this model, ancient protein structures were formed by self-assembling aggregates of short polypeptides. Subsequently, and perhaps concomitantly with the evolution of higher fidelity DNA replication and repair systems, single polypeptide domains arose from the fusion of ADSs genes. Thus modern protein domains may have a polyphyletic origin [165].

The four possible evolutionary scenarios for the evolution of sequence- and structure-similar protein fragments in non-homologous protein folds [165].

- Under functional constraints nature could have invented the sequence- and structure-similar protein fragments more than once by **convergence**. However, there is one problem with this evolutionary route. Many proteins have similar functions with no molecular similarities at all. It is very likely that different folds, once evolved, have converged to similar functions by more conventional evolutionary events, such as point mutations [165].

- A second possibility is that sequence- and structure-similar protein fragments present in different folds, have arisen by **divergence**. After descending from their ancestors the divergent evolutionary mechanisms, such as permutations, deletions, insertions, and rearrangements, might account for the drastic change in the variable regions of the folds while retaining the core of their ancestors. This might have resulted in common sequence- and structure-similar protein fragments with different folds. However, there are no obvious divergent evolutionary mechanisms that might account for many of the observed fold differences [165].

- A third evolutionary mechanism is thought to be due to partial duplication of one gene and subsequent recombination within a domain-free region of a second gene. In many instances, partially duplicated gene might have been inserted into another domain by recombination within a domain encoding region of the second gene [165].

- A fourth possibility explains the evolutionary origin of domain itself. In this scenario, the single-chain contemporary domains arose from the fusion of more ancient mini-genes that encoded sequence- and structure-similar protein fragments. Therefore many of the core folds observed today may contain homologous building blocks. Peptides forming these building blocks

> might not have in themselves, the ability to fold, but might have emerged as cofactors supporting RNA-based replication and catalysis ("RNA world"). Their association into larger structures and eventual fusion into polypeptide chains would have allowed them to become independent of their RNA scaffold, leading to the evolution of a novel type of macromolecule: the folded protein [79].

In the protein world, living fossils are the molecules whose role in cellular processes is so central that further modification has become nearly impossible. They are essentially frozen in time. Ribosome is the central component in protein synthesis; it emerged early in the evolution of life and was essentially fixed at the time of the last common ancestor. Correspondingly, the core complement of ribosomal proteins is more than 40 % identical between all living organisms. The crystal structures show that only few ribosomal proteins are fully folded; many have folded domains 'sprouting' from a part of the polypeptide chain which lacks secondary structure and is tightly associated with the RNA, while some proteins have neither folded structure nor any secondary structure. Cumulatively, this scenario is one of the progressive structural emancipation, from complete dependence on the RNA template to nearly a full independent snapshot of the time when proteins learned to fold [79].

We have discovered the sequence- and structure-similar protein fragments in the ribosomal protein S20 (RPS20) from *Thermus thermophilus* that resembles the repeating units of TPRs and attempted to build TPR-like domain by repetition of this fragment. Our results provides the evidence that hints at the evolutionary origin of domain by constructing new TPR-like domains from the sequence- and structure-similar α-hairpins from different folds. We have shown that computational methods can reliably detect the remnants of ancient peptides, which have been used by nature in building different folds. Experimentally, we have demonstrated the role of repetition in the evolution of novel protein domains by simply duplicating the α-hairpin fragments which are similar in both structure and sequence to the TPR repeating unit. These proteins appear to have stable structures, and have features typical to the naturally occurring proteins (Figure 3.23, Figure 3.25 and 3.27). Tom20 and RPS20 α-hairpin derived proteins have temperature melts typical to the tightly packed proteins as illustrated in Figure 3.25.

Table 4.1: **Summary of experimental results.** The cell in the Table where the particular experiment cannot be carried out or does not make any sense are empty.

Parent	**Proteins**	**CD spectra**	**Thermal Denaturation**	**Fluorescence Spectra**	**Verdict**
Tom20	Tom20-1	α-helical like	Cooperative		Native-like
	Tom20-2	α-helical like	Cooperative		Native-like
	Tom20-3	α-helical like	Cooperative		Native-like
RPS20	RPS20-1	α-helical like	Cooperative		Native-like
	RPS20-2	α-helical like	Cooperative		Native-like
	RPS20-3	α-helical like	Cooperative		Native-like
PLC	PLC-1	Anomalous	Non-coopearive		Unfolded
	PLC-2	Anomalous	Non-coopearive	Shift	Unfolded
	PLC-3	Anomalous	Non-coopearive	Shift	Unfolded
HSC	HSC-1	Random-coil			Unfolded
	HSC-2	Random-coil			Unfolded
	HSC-3	Random-coil		Shift	Unfolded
BGA	BGA-1	Random-coil			Unfolded
	BGA-2	Random-coil			Unfolded
	BGA-3	Random-coil			Unfolded

The solution properties of the designed proteins are summarised in Table 4.1, which are directly correlating with the statistical significance of their parent α-hairpins likely to be TPR repeating unit (Table 3.5). Six proteins (Tom20-1/2/3 and RPS20-1/2/3) behave like natural proteins, whereas other proteins (PLC-1/2/3, HSC-1/2/3 and BGA-1/2/3) completely behave like random-coils. In each of these cases where the proteins behaviour is more like natural proteins (Tom20-1/2/3 and RPS20-1/2/3), the parent domains are mainly made up of an α-hairpin with an extra "solvating" helix (Figure 3.8). Minimal characteristics feature of these domains might be advantageous to evolve as a new TPR-like domain by duplication. These results are consistent with the notion that evolution drives from simplicity to complexity [82]. More detailed comparisons of these proteins to the naturally occurring TPR proteins must await high-resolution determination of the structure for one of these proteins. In contract, the remaining three α-hairpins were already a part of a complex protein structures (Figure 3.9) therefore, it is less probable that they would evolve into new TPR-like domains.

The results described here establish that new domains can be constructed from the protein fragments that are similar to a repeating unit of a target domain. These encouraging results suggest that it is possible to make a domain of interest from the set of smaller structural motifs.

References

[1] Pace NR: **Origin of life–facing up to the physical setting**. *Cell* 1991, **65**:531–533.

[2] Lazcano A, Miller SL: **The origin and early evolution of life: prebiotic chemistry, the pre-RNA world, and time**. *Cell* 1996, **85**:793–798.

[3] Miller SL, Bada JL: **Submarine hot springs and the origin of life**. *Nature* 1988, **334**:609–611.

[4] Lazcano A, Miller SL: **How long did it take for life to begin and evolve to cyanobacteria?** *J Mol Evol* 1994, **39**:546–554.

[5] Joyce GF: **The antiquity of RNA-based evolution**. *Nature* 2002, **418**:214–221.

[6] Kasting JF: **Earth's early atmosphere**. *Science* 1993, **259**:920–926.

[7] Evans D, Marquez SM, Pace NR: **RNase P: interface of the RNA and protein worlds**. *Trends Biochem Sci* 2006, **31**:333–341.

[8] Shapiro R: **Prebiotic ribose synthesis: a critical analysis**. *Orig Life Evol Biosph* 1988, **18**:71–85.

[9] Ban N, Nissen P, Hansen J, Moore PB, Steitz TA: **The complete atomic structure of the large ribosomal subunit at 2.4 A resolution**. *Science* 2000, **289**:905–920.

[10] Wimberly BT, Brodersen DE, Clemons WMJ, Morgan-Warren RJ, Carter AP, Vonrhein C, Hartsch T, Ramakrishnan V: **Structure of the 30S ribosomal subunit**. *Nature* 2000, **407**:327–339.

[11] Yusupov MM, Yusupova GZ, Baucom A, Lieberman K, Earnest TN, Cate JH, Noller HF: **Crystal structure of the ribosome at 5.5 A resolution**. *Science* 2001, **292**:883–896.

[12] Ramakrishnan V, White SW: **Ribosomal protein structures: insights into the architecture, machinery and evolution of the ribosome**. *Trends Biochem Sci* 1998, **23**:208–212.

[13] Doudna JA, Cech TR: **The chemical repertoire of natural ribozymes**. *Nature* 2002, **418**:222–228.

[14] Benner SA, Ellington AD, Tauer A: **Modern metabolism as a palimpsest of the RNA world**. *Proc Natl Acad Sci U S A* 1989, **86**:7054–7058.

[15] Gilbert W, Marchionni M, McKnight G: **On the antiquity of introns**. *Cell* 1986, **46**:151–153.

[16] Poole AM, Logan DT: **Modern mRNA proofreading and repair: clues that the last universal common ancestor possessed an RNA genome?** *Mol Biol Evol* 2005, **22**:1444–1455.

[17] Roth A, Breaker RR: **An amino acid as a cofactor for a catalytic polynucleotide**. *Proc Natl Acad Sci U S A* 1998, **95**:6027–6031.

[18] Joyce GF: **Nucleic acid enzymes: playing with a fuller deck**. *Proc Natl Acad Sci U S A* 1998, **95**:5845–5847.

[19] Orgel LE: **The origin of polynucleotide-directed protein synthesis**. *J Mol Evol* 1989, **29**:465–474.

[20] Orgel LE: **Evolution of the genetic apparatus**. *J Mol Biol* 1968, **38**:381–393.

[21] B SJ, L DA: **Crystalline catalase**. *J Biol Chem* 1937, **121**:417–424.

[22] Frieden E: **The chemical elements of life**. *Sci Am* 1972, **227**:52–60.

[23] Lee D, Grant A, Buchan D, Orengo C: **A structural perspective on genome evolution**. *Curr Opin Struct Biol* 2003, **13**:359–369.

[24] Brewster JH, Laskowski MJ: **Left-handed comments**. *Science* 1992, **258**:1289; author reply 1290.

[25] Sanger F: **Sequences, sequences, and sequences**. *Annu Rev Biochem* 1988, **57**:1–28.

[26] PAULING L, COREY RB: **Atomic coordinates and structure factors for two helical configurations of polypeptide chains**. *Proc Natl Acad Sci U S A* 1951, **37**:235–240.

[27] Venkatachalam CM, Ramachandran GN: **Conformation of polypeptide chains**. *Annu Rev Biochem* 1969, **38**:45–82.

[28] Sanchez GR, Chaiken IM, Anfinsen CB: **Structure-function relationships at the active site of nuclease-T'**. *J Biol Chem* 1973, **248**:3653–3659.

[29] Chan HS, Dill KA: **Polymer principles in protein structure and stability**. *Annu Rev Biophys Biophys Chem* 1991, **20**:447–490.

[30] Anfinsen CB: **The formation and stabilization of protein structure**. *Biochem J* 1972, **128**:737–749.

[31] Kato I, Anfinsen CB: **On the stabilization of ribonuclease S-protein by ribonuclease S-peptide**. *J Biol Chem* 1969, **244**:1004–1007.

[32] Epstein HF, Schechter AN, Chen RF, Anfinsen CB: **Folding of staphylococcal nuclease: kinetic studies of two processes in acid renaturation**. *J Mol Biol* 1971, **60**:499–508.

[33] Anfinsen CB: **Principles that govern the folding of protein chains**. *Science* 1973, **181**:223–230.

[34] Ellis RJ, Hartl FU: **Principles of protein folding in the cellular environment**. *Curr Opin Struct Biol* 1999, **9**:102–110.

[35] Rothman JE: **Polypeptide chain binding proteins: catalysts of protein folding and related processes in cells**. *Cell* 1989, **59**:591–601.

[36] Van Dyk TK, Gatenby AA, LaRossa RA: **Demonstration by genetic suppression of interaction of GroE products with many proteins**. *Nature* 1989, **342**:451–453.

[37] Ellis RJ: **Protein folding: inside the cage**. *Nature* 2006, **442**:360–362.

[38] Creighton TE: **How important is the molten globule for correct protein folding?** *Trends Biochem Sci* 1997, **22**:6–10.

[39] Ptitsyn OB: **How the molten globule became**. *Trends Biochem Sci* 1995, **20**:376–379.

[40] Dill KA, Chan HS: **From Levinthal to pathways to funnels**. *Nat Struct Biol* 1997, **4**:10–19.

[41] Schechter AN, Chen RF, Anfinsen CB: **Kinetics of folding of staphylococcal nuclease**. *Science* 1970, **167**:886–887.

[42] Cremades N, Sancho J, Freire E: **The native-state ensemble of proteins provides clues for folding, misfolding and function**. *Trends Biochem Sci* 2006, **31**:494–496.

[43] James LC, Tawfik DS: **Conformational diversity and protein evolution–a 60-year-old hypothesis revisited**. *Trends Biochem Sci* 2003, **28**:361–368.

[44] Ellis RJ, Hartl FU: **Protein folding in the cell: competing models of chaperonin function**. *FASEB J* 1996, **10**:20–26.

[45] Clark PL: **Protein folding in the cell: reshaping the folding funnel**. *Trends Biochem Sci* 2004, **29**:527–534.

[46] Bork P, Koonin EV: **Predicting functions from protein sequences–where are the bottlenecks?** *Nat Genet* 1998, **18**:313–318.

[47] Needleman SB, Wunsch CD: **A general method applicable to the search for similarities in the amino acid sequence of two proteins**. *J Mol Biol* 1970, **48**:443–453.

[48] Pearson WR, Lipman DJ: **Improved tools for biological sequence comparison**. *Proc Natl Acad Sci U S A* 1988, **85**:2444–2448.

[49] Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: **Basic local alignment search tool**. *J Mol Biol* 1990, **215**:403–410.

[50] Smith TF, Waterman MS: **Identification of common molecular subsequences**. *J Mol Biol* 1981, **147**:195–197.

[51] Brenner SE, Chothia C, Hubbard TJ: **Assessing sequence comparison methods with reliable structurally identified distant evolutionary relationships**. *Proc Natl Acad Sci U S A* 1998, **95**:6073–6078.

[52] Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs**. *Nucleic Acids Res* 1997, **25**:3389–3402.

[53] Park J, Teichmann SA, Hubbard T, Chothia C: **Intermediate sequences increase the detection of homology between sequences**. *J Mol Biol* 1997, **273**:349–354.

[54] Park J, Karplus K, Barrett C, Hughey R, Haussler D, Hubbard T, Chothia C: **Sequence comparisons using multiple sequences detect three times as many remote homologues as pairwise methods**. *J Mol Biol* 1998, **284**:1201–1210.

[55] Rychlewski L, Jaroszewski L, Li W, Godzik A: **Comparison of sequence profiles. Strategies for structural predictions using sequence information**. *Protein Sci* 2000, **9**:232–241.

[56] Pietrokovski S: **Searching databases of conserved sequence regions by aligning protein multiple-alignments**. *Nucleic Acids Res* 1996, **24**:3836–3845.

[57] Yona G, Levitt M: **Within the twilight zone: a sensitive profile-profile comparison tool based on information theory**. *J Mol Biol* 2002, **315**:1257–1275.

[58] Sadreyev R, Grishin N: **COMPASS: a tool for comparison of multiple protein alignments with assessment of statistical significance**. *J Mol Biol* 2003, **326**:317–336.

[59] Soding J: **Protein homology detection by HMM-HMM comparison**. *Bioinformatics* 2005, **21**:951–960.

[60] Kinch LN, Grishin NV: **Evolution of protein structures and functions**. *Curr Opin Struct Biol* 2002, **12**:400–408.

[61] Kolodny R, Petrey D, Honig B: **Protein structure comparison: implications for the nature of 'fold space', and structure and function prediction**. *Curr Opin Struct Biol* 2006, **16**:393–398.

[62] Muirhead H, Cox JM, Mazzarella L, Perutz MF: **Structure and function of haemoglobin. 3. A three-dimensional fourier synthesis of human deoxyhaemoglobin at 5.5 Angstrom resolution**. *J Mol Biol* 1967, **28**:117–156.

[63] Drenth J, Hol WG, Jansonius JN, Koekoek R: **Subtilisin Novo. The three-dimensional structure and its comparison with subtilisin BPN'**. *Eur J Biochem* 1972, **26**:177–181.

[64] Huber R, Epp O, Steigemann W, Formanek H: **The atomic structure of erythrocruorin in the light of the chemical sequence and its comparison with myoglobin**. *Eur J Biochem* 1971, **19**:42–50.

[65] Rao ST, Rossmann MG: **Comparison of super-secondary structures in proteins**. *J Mol Biol* 1973, **76**:241–256.

[66] Holm L, Sander C: **Protein structure comparison by alignment of distance matrices**. *J Mol Biol* 1993, **233**:123–138.

[67] Shindyalov IN, Bourne PE: **Protein structure alignment by incremental combinatorial extension (CE) of the optimal path**. *Protein Eng* 1998, **11**:739–747.

[68] Taylor WR, Orengo CA: **Protein structure alignment**. *J Mol Biol* 1989, **208**:1–22.

[69] Yang AS, Honig B: **An integrated approach to the analysis and modeling of protein sequences and structures. I. Protein structural alignment and a quantitative measure for protein structural distance**. *J Mol Biol* 2000, **301**:665–678.

[70] Krissinel E, Henrick K: **Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions**. *Acta Crystallogr D Biol Crystallogr* 2004, **60**:2256–2268.

[71] Schonbrun J, Wedemeyer WJ, Baker D: **Protein structure prediction in 2002**. *Curr Opin Struct Biol* 2002, **12**:348–354.

[72] Ponder JW, Richards FM: **Tertiary templates for proteins. Use of packing criteria in the enumeration of allowed sequences for different structural classes**. *J Mol Biol* 1987, **193**:775–791.

[73] 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.

[74] Yue K, Dill KA: **Inverse protein folding problem: designing polymer sequences**. *Proc Natl Acad Sci U S A* 1992, **89**:4163–4167.

[75] Orengo CA, Jones DT, Thornton JM: **Protein superfamilies and domain superfolds**. *Nature* 1994, **372**:631–634.

[76] Jones DT: **Structural biology. Learning to speak the language of proteins**. *Science* 2003, **302**:1347–1348.

[77] Lang D, Thoma R, Henn-Sax M, Sterner R, Wilmanns M: **Structural evidence for evolution of the beta/alpha barrel scaffold by gene duplication and fusion**. *Science* 2000, **289**:1546–1550.

[78] Hocker B, Beismann-Driemeyer S, Hettwer S, Lustig A, Sterner R: **Dissection of a (betaalpha)8-barrel enzyme into two folded halves**. *Nat Struct Biol* 2001, **8**:32–36.

[79] Soding J, Lupas AN: **More than the sum of their parts: on the evolution of proteins from peptides**. *Bioessays* 2003, **25**:837–846.

[80] Kondrashov AS: **Deleterious mutations and the evolution of sexual reproduction**. *Nature* 1988, **336**:435–440.

[81] Jordan IK, Kondrashov FA, Adzhubei IA, Wolf YI, Koonin EV, Kondrashov AS, Sunyaev S: **A universal trend of amino acid gain and loss in protein evolution**. *Nature* 2005, **433**:633–638.

[82] Ohta T: **Role of gene duplication in evolution**. *Genome* 1989, **31**:304–310.

[83] Vision TJ, Brown DG, Tanksley SD: **The origins of genomic duplications in Arabidopsis**. *Science* 2000, **290**:2114–2117.

[84] McLysaght A, Hokamp K, Wolfe KH: **Extensive genomic duplication during early chordate evolution**. *Nat Genet* 2002, **31**:200–204.

[85] Wolfe KH, Shields DC: **Molecular evidence for an ancient duplication of the entire yeast genome**. *Nature* 1997, **387**:708–713.

[86] Jackson EN, Yanofsky C: **Duplication-translocations of tryptophan operon genes in Escherichia coli**. *J Bacteriol* 1973, **116**:33–40.

[87] Lynch M, Conery JS: **The evolutionary fate and consequences of duplicate genes**. *Science* 2000, **290**:1151–1155.

[88] Marcotte EM, Pellegrini M, Yeates TO, Eisenberg D: **A census of protein repeats**. *J Mol Biol* 1999, **293**:151–160.

[89] Ponting CP, Russell RB: **Swaposins: circular permutations within genes encoding saposin homologues**. *Trends Biochem Sci* 1995, **20**:179–180.

[90] Lindqvist Y, Schneider G: **Circular permutations of natural protein sequences: structural evidence**. *Curr Opin Struct Biol* 1997, **7**:422–427.

[91] Doolittle RF: **The multiplicity of domains in proteins**. *Annu Rev Biochem* 1995, **64**:287–314.

[92] Kajava AV: **Review: proteins with repeated sequence–structural prediction and modeling**. *J Struct Biol* 2001, **134**:132–144.

[93] Doolittle WF: **Phylogenetic classification and the universal tree**. *Science* 1999, **284**:2124–2129.

[94] Chen L, DeVries AL, Cheng CH: **Convergent evolution of antifreeze glycoproteins in Antarctic notothenioid fish and Arctic cod**. *Proc Natl Acad Sci U S A* 1997, **94**:3817–3822.

[95] Chen L, DeVries AL, Cheng CH: **Evolution of antifreeze glycoprotein gene from a trypsinogen gene in Antarctic notothenioid fish**. *Proc Natl Acad Sci U S A* 1997, **94**:3811–3816.

[96] Kohl A, Binz HK, Forrer P, Stumpp MT, Pluckthun A, Grutter MG: **Designed to be stable: crystal structure of a consensus ankyrin repeat protein**. *Proc Natl Acad Sci U S A* 2003, **100**:1700–1705.

[97] Main ERG, Lowe AR, Mochrie SGJ, Jackson SE, Regan L: **A recurring theme in protein engineering: the design, stability and folding of repeat proteins**. *Curr Opin Struct Biol* 2005, **15**:464–471.

[98] Groves MR, Barford D: **Topological characteristics of helical repeat proteins**. *Curr Opin Struct Biol* 1999, **9**:383–389.

[99] Kobe B, Kajava AV: **When protein folding is simplified to protein coiling: the continuum of solenoid protein structures**. *Trends Biochem Sci* 2000, **25**:509–515.

[100] Main ERG, Jackson SE, Regan L: **The folding and design of repeat proteins: reaching a consensus**. *Curr Opin Struct Biol* 2003, **13**:482–489.

[101] Sikorski RS, Boguski MS, Goebl M, Hieter P: **A repeating amino acid motif in CDC23 defines a family of proteins and a new relationship among genes required for mitosis and RNA synthesis**. *Cell* 1990, **60**:307–317.

[102] D'Andrea LD, Regan L: **TPR proteins: the versatile helix**. *Trends Biochem Sci* 2003, **28**:655–662.

[103] Lamb JR, Tugendreich S, Hieter P: **Tetratrico peptide repeat interactions: to TPR or not to TPR?** *Trends Biochem Sci* 1995, **20**:257–259.

[104] Kyrpides NC, Woese CR: **Tetratrico-peptide-repeat proteins in the archaeon Methanococcus jannaschii**. *Trends Biochem Sci* 1998, **23**:245–247.

[105] Das AK, Cohen PW, Barford D: **The structure of the tetratricopeptide repeats of protein phosphatase 5: implications for TPR-mediated protein-protein interactions**. *EMBO J* 1998, **17**:1192–1199.

[106] Forrer P, Binz HK, Stumpp MT, Pluckthun A: **Consensus design of repeat proteins**. *Chembiochem* 2004, **5**:183–189.

[107] Magliery TJ, Regan L: **Beyond consensus: statistical free energies reveal hidden interactions in the design of a TPR motif**. *J Mol Biol* 2004, **343**:731–745.

[108] Small ID, Peeters N: **The PPR motif - a TPR-related motif prevalent in plant organellar proteins**. *Trends Biochem Sci* 2000, **25**:46–47.

[109] Grant B, Greenwald I: **The Caenorhabditis elegans sel-1 gene, a negative regulator of lin-12 and glp-1, encodes a predicted extracellular protein**. *Genetics* 1996, **143**:237–247.

[110] Kotera E, Tasaka M, Shikanai T: **A pentatricopeptide repeat protein is essential for RNA editing in chloroplasts**. *Nature* 2005, **433**:326–330.

[111] Scheufler C, Brinker A, Bourenkov G, Pegoraro S, Moroder L, Bartunik H, Hartl FU, Moarefi I: **Structure of TPR domain-peptide complexes: critical elements in the assembly of the Hsp70-Hsp90 multichaperone machine**. *Cell* 2000, **101**:199–210.

[112] Sonnhammer EL, Eddy SR, Durbin R: **Pfam: a comprehensive database of protein domain families based on seed alignments**. *Proteins* 1997, **28**:405–420.

[113] Schultz J, Milpetz F, Bork P, Ponting CP: **SMART, a simple modular architecture research tool: identification of signaling domains**. *Proc Natl Acad Sci U S A* 1998, **95**:5857–5864.

[114] Andrade MA, Ponting CP, Gibson TJ, Bork P: **Homology-based method for identification of protein repeats using statistical significance estimates**. *J Mol Biol* 2000, **298**:521–537.

[115] Steegborn C, Danot O, Huber R, Clausen T: **Crystal structure of transcription factor MalT domain III: a novel helix repeat fold implicated in regulated oligomerization**. *Structure (Camb)* 2001, **9**:1051–1060.

[116] Dohm JA, Lee SJ, Hardwick JM, Hill RB, Gittis AG: **Cytosolic domain of the human mitochondrial fission protein fis1 adopts a TPR fold**. *Proteins* 2004, **54**:153–156.

[117] Boos W, Shuman H: **Maltose/maltodextrin system of Escherichia coli: transport, metabolism, and regulation**. *Microbiol Mol Biol Rev* 1998, **62**:204–229.

[118] Karpenahalli MR, Soeding J, Lupas AN: **TPRpred: a tool for prediction of TPR, PPR and SEL1-like repeats from protein sequences**. *BMC Bioinformatics* 2006. in press.

[119] Zhang H, Grishin NV: **The alpha-subunit of protein prenyltransferases is a member of the tetratricopeptide repeat family**. *Protein Sci* 1999, **8**:1658–1667.

[120] Park HS, Nam SH, Lee JK, Yoon CN, Mannervik B, Benkovic SJ, Kim HS: **Design and evolution of new catalytic activity with an existing protein scaffold**. *Science* 2006, **311**:535–538.

[121] Andrade MA, Perez-Iratxeta C, Ponting CP: **Protein repeats: structures, functions, and evolution**. *J Struct Biol* 2001, **134**:117–131.

[122] Perry AJ, Hulett JM, Likic VA, Lithgow T, Gooley PR: **Convergent evolution of receptors for protein import into mitochondria**. *Curr Biol* 2006, **16**:221–229.

[123] Main ERG, Xiong Y, Cocco MJ, D'Andrea L, Regan L: **Design of stable alpha-helical arrays from an idealized TPR motif**. *Structure* 2003, **11**:497–508.

[124] Cortajarena AL, Kajander T, Pan W, Cocco MJ, Regan L: **Protein design to understand peptide ligand recognition by tetratricopeptide repeat proteins**. *Protein Eng Des Sel* 2004, **17**:399–409.

[125] Main ERG, Stott K, Jackson SE, Regan L: **Local and long-range stability in tandemly arrayed tetratricopeptide repeats**. *Proc Natl Acad Sci U S A* 2005, **102**:5721–5726.

[126] Abe Y, Shodai T, Muto T, Mihara K, Torii H, Nishikawa S, Endo T, Kohda D: **Structural basis of presequence recognition by the mitochondrial protein import receptor Tom20**. *Cell* 2000, **100**:551–560.

[127] Murzin AG, Brenner SE, Hubbard T, Chothia C: **SCOP: a structural classification of proteins database for the investigation of sequences and structures**. *J Mol Biol* 1995, **247**:536–540.

[128] **HMMER: profile HMMs for protein sequence analysis**[http://hmmer.wustl.edu/].

[129] Bernstein FC, Koetzle TF, Williams GJ, Meyer EFJ, Brice MD, Rodgers JR, Kennard O, Shimanouchi T, Tasumi M: **The Protein Data Bank: a computer-based archival file for macromolecular structures**. *J Mol Biol* 1977, **112**:535–542.

[130] Guex N, Peitsch MC: **SWISS-MODEL and the Swiss-PdbViewer: an environment for comparative protein modeling**. *Electrophoresis* 1997, **18**:2714–2723.

[131] Schuster-Bockler B, Schultz J, Rahmann S: **HMM Logos for visualization of protein families**. *BMC Bioinformatics* 2004, **5**:7.

[132] Russell RB, Barton GJ: **Multiple protein sequence alignment from tertiary structure comparison: assignment of global and residue confidence levels**. *Proteins* 1992, **14**:309–323.

[133] Godzik A, Skolnick J, Kolinski A: **Regularities in interaction patterns of globular proteins**. *Protein Eng* 1993, **6**:801–810.

[134] Nakamura Y, Gojobori T, Ikemura T: **Codon usage tabulated from international DNA sequence databases: status for the year 2000**. *Nucleic Acids Res* 2000, **28**:292.

[135] Sali A, Blundell TL: **Comparative protein modelling by satisfaction of spatial restraints**. *J Mol Biol* 1993, **234**:779–815.

[136] Schagger H, von Jagow G: **Tricine-sodium dodecyl sulfate-polyacrylamide gel electrophoresis for the separation of proteins in the range from 1 to 100 kDa**. *Anal Biochem* 1987, **166**:368–379.

[137] Clamp M, Cuff J, Searle SM, Barton GJ: **The Jalview Java alignment editor**. *Bioinformatics* 2004, **20**:426–427.

[138] Li W, Jaroszewski L, Godzik A: **Clustering of highly homologous sequences to reduce the size of large protein databases**. *Bioinformatics* 2001, **17**:282–283.

[139] Li W, Jaroszewski L, Godzik A: **Tolerating some redundancy significantly speeds up clustering of large protein databases**. *Bioinformatics* 2002, **18**:77–82.

[140] De Schrijver A, De Mot R: **A subfamily of MalT-related ATP-dependent regulators in the LuxR family**. *Microbiology* 1999, **145 (Pt 6)**:1287–1288.

[141] Leipe DD, Koonin EV, Aravind L: **STAND, a class of P-loop NTPases including animal and plant regulators of programmed cell death: multiple, complex domain architectures, unusual phyletic patterns, and evolution by horizontal gene transfer**. *J Mol Biol* 2004, **343**:1–28.

[142] Ammelburg M, Frickey T, Lupas A: **Classification of AAA+ proteins**. *J Struct Biol* 2006. in press.

[143] Chandonia JM, Hon G, Walker NS, Lo Conte L, Koehl P, Levitt M, Brenner SE: **The ASTRAL Compendium in 2004**. *Nucleic Acids Res* 2004, **32**:189–192.

[144] Sadreyev RI, Baker D, Grishin NV: **Profile-profile comparisons by COMPASS predict intricate homologies between protein families**. *Protein Sci* 2003, **12**:2262–2272.

[145] Kraulis PJ: **MOLSCRIPT: a program to produce both detailed and schematic plots of protein structures**. *J App Cryst* 1991, **24**:946–950.

[146] Merritt EA, Murphy ME: **Raster3D Version 2.0. A program for photorealistic molecular graphics**. *Acta Crystallogr D Biol Crystallogr* 1994, **50**:869–873.

[147] Salem GM, Hutchinson EG, Orengo CA, Thornton JM: **Correlation of observed fold frequency with the occurrence of local structural motifs**. *J Mol Biol* 1999, **287**:969–981.

[148] Kumar S, Bansal M: **Dissecting alpha-helices: position-specific analysis of alpha-helices in globular proteins**. *Proteins* 1998, **31**:460–476.

[149] Aurora R, Rose GD: **Helix capping**. *Protein Sci* 1998, **7**:21–38.

[150] Dasgupta S, Bell JA: **Design of helix ends. Amino acid preferences, hydrogen bonding and electrostatic interactions**. *Int J Pept Protein Res* 1993, **41**:499–511.

[151] Richardson JS, Richardson DC: **Amino acid preferences for specific locations at the ends of alpha helices**. *Science* 1988, **240**:1648–1652.

[152] Dyall SD, Brown MT, Johnson PJ: **Ancient invasions: from endosymbionts to organelles**. *Science* 2004, **304**:253–257.

[153] Gray MW, Burger G, Lang BF: **Mitochondrial evolution**. *Science* 1999, **283**:1476–1481.

[154] Pfanner N, Geissler A: **Versatility of the mitochondrial protein import machinery**. *Nat Rev Mol Cell Biol* 2001, **2**:339–349.

[155] Lister R, Whelan J: **Mitochondrial protein import: convergent solutions for receptor structure**. *Curr Biol* 2006, **16**:197–199.

[156] Neupert W: **Protein import into mitochondria**. *Annu Rev Biochem* 1997, **66**:863–917.

[157] Justin N, Walker N, Bullifent HL, Songer G, Bueschel DM, Jost H, Naylor C, Miller J, Moss DS, Titball RW, Basak AK: **The first strain of Clostridium perfringens isolated from an avian source has an alpha-toxin with divergent structural and kinetic properties**. *Biochemistry* 2002, **41**:6253–6262.

[158] Cupp-Vickery JR, Vickery LE: **Crystal structure of Hsc20, a J-type Co-chaperone from Escherichia coli**. *J Mol Biol* 2000, **304**:835–845.

[159] Aleshin AE, Feng PH, Honzatko RB, Reilly PJ: **Crystal structure and evolution of a prokaryotic glucoamylase**. *J Mol Biol* 2003, **327**:61–73.

[160] Goff SA, Goldberg AL: **Production of abnormal proteins in E. coli stimulates transcription of lon and other heat shock genes**. *Cell* 1985, **41**:587–595.

[161] Kandror O, Busconi L, Sherman M, Goldberg AL: **Rapid degradation of an abnormal protein in Escherichia coli involves the chaperones GroEL and GroES**. *J Biol Chem* 1994, **269**:23575–23582.

[162] Drew D, Sjostrand D, Nilsson J, Urbig T, Chin Cn, de Gier JW, von Heijne G: **Rapid topology mapping of Escherichia coli inner-membrane proteins by prediction and PhoA/GFP fusion analysis**. *Proc Natl Acad Sci U S A* 2002, **99**:2690–2695.

[163] Wang H, Chong S: **Visualization of coupled protein folding and binding in bacteria and purification of the heterodimeric complex**. *Proc Natl Acad Sci U S A* 2003, **100**:478–483.

[164] Yano M, Terada K, Mori M: **Mitochondrial import receptors Tom20 and Tom22 have chaperone-like activity**. *J Biol Chem* 2004, **279**:10808–10813.

[165] Lupas AN, Ponting CP, Russell RB: **On the evolution of protein folds: are similar motifs in different protein folds the result of convergence, insertion, or relics of an ancient peptide world?** *J Struct Biol* 2001, **134**:191–203.

[166] Minor DLJ, Kim PS: **Context-dependent secondary structure formation of a designed protein sequence**. *Nature* 1996, **380**:730–734.

[167] Tidow H, Lauber T, Vitzithum K, Sommerhoff CP, Rosch P, Marx UC: **The solution structure of a chimeric LEKTI domain reveals a chameleon sequence**. *Biochemistry* 2004, **43**:11238–11247.

[168] Andreeva A, Murzin AG: **Evolution of protein fold in the presence of functional constraints**. *Curr Opin Struct Biol* 2006, **16**:399–408.

[169] Benner SA, Cohen MA, Gonnet GH: **Empirical and structural models for insertions and deletions in the divergent evolution of proteins**. *J Mol Biol* 1993, **229**:1065–1082.

[170] Pascarella S, Argos P: **Analysis of insertions/deletions in protein structures**. *J Mol Biol* 1992, **224**:461–471.

[171] Grishin NV: **Fold change in evolution of protein structures**. *J Struct Biol* 2001, **134**:167–185.

[172] Russell RB, Ponting CP: **Protein fold irregularities that hinder sequence analysis**. *Curr Opin Struct Biol* 1998, **8**:364–371.

CURRICULUM VITAE

PERSONAL DETAILS

Name	**Manjunatha Karpenahalli Ranganathappa**
Date of birth	May, 4th, 1975
Place of birth	Karpenahalli, Karnataka State, India
Nationality	Indian

EDUCATION

2002 - 2006	**PhD**, Max-Planck-Institute for Developmental Biology, Tübingen, Germany
2000 - 2001	**Advanced Post Graduate Diploma in Bioinformatics**, Madurai Kamaraj University, Madurai, India
1997 - 1999	**Master of Science in Biochemistry**, Bangalore University, Bangalore, India
1994 - 1997	**Bachelor of Science in Chemistry, Botany and Zoology**, Bangalore University, Tumkur, India

PUBLICATIONS

1. **Karpenahalli, M.R.**, Lupas, A.N. and Söding, J. "TPRpred: a tool for prediction of TPR-, PPR- and SEL1-like repeats from protein sequences" [*In press, BMC Bioinformatics.*]

2. **Karpenahalli, M.R.**, Martin, J. and Lupas, A.N. "Exploring the evolutionary origin of protein domain by designing new TPR-like domains from the α-hairpin protein fragments" [*Submitted.*]