On the Evolution of Proteins from Peptides

Dissertation
der Mathematisch-Naturwissenschaftlichen Fakultät
der Eberhard Karls Universität Tübingen
zur Erlangung des Grades eines
Doktors der Naturwissenschaften
(Dr. rer. nat.)

vorgelegt von
Vikram Alva Kullanja
aus Bantwal (Indien)

Tübingen
2012

Tag der mündlichen Qualifikation:	27.09.2012
Dekan:	Prof. Dr. Wolfgang Rosenstiel
1. Berichterstatter:	Prof. Dr. Andrei N. Lupas
2. Berichterstatter:	Prof. Dr. Oliver Kohlbacher

Erklärung

Hiermit erkläre ich, dass ich die Arbeit selbständig und nur mit den angegebenen Hilfsmitteln angefertigt habe und dass alle Stellen, die im Wortlaut oder dem Sinne nach anderen Werken entnommen sind, durch Angaben der Quellen kenntlich gemacht sind.

Tübingen, März 2012 *Vikram Alva Kullanja*

This thesis is dedicated to the memory of my father, K. Jagannatha Alva.

Abstract

Though seemingly endless, the diversity of proteins in nature is in fact narrowly confined. Many proteins share recognizable similarity in sequence and structure, since they arose by amplification, recombination, and divergence from a basic complement of autonomously folding modules, referred to as domains, many of which date back to the time of the Last Universal Common Ancestor. Indeed, sequence comparison of modern proteins shows that they fall into only about 10,000 domain families, which can be further grouped into just about 3000 broader evolutionary superfamilies. Beyond this, superfamilies are assigned to one of about 1000 folds based on the topological arrangement of their secondary structural elements. The prevailing view holds that folds are analogous in character, the similarity between different superfamilies of one fold being the result of convergent evolution. However, the recent growth of molecular databases and advances in sequence comparison methods have led to the discovery of many distant evolutionary relationships that transcend the boundaries of superfamilies, showing that not all of them arose independently. The first aim of this thesis was to determine how widespread such distant relationships are. To this end, I clustered domains representative of known fold types by their sequence similarity, a property that reflects common descent. The obtained cluster map shows that while some highly populated folds indeed appear to have evolved convergently, most domains of the same fold arose from an ancestral prototype, revealing that proteins are much less polyphyletic than previously assumed.

Whereas it is widely accepted that modern proteins arose by combinatorial shuffling of a limited set of domains, the origin of this set itself is poorly understood. Even the simplest domains are too complex to have arisen *de novo*. If so, how did the first domains emerge? This question formed the second aim of this thesis. One theory for the origin of domains, the antecedent domain segment theory, proposes that they themselves arose from an even smaller pool of peptides with secondary structure propensity, which emerged as cofactors in the RNA world. Progressively more stable domains evolved from this set by amplification and by accretion, that is, by additive assemblage of simple structural elements. If this is true, many modern domains might still contain vestiges of the ancient peptides they arose from.

To investigate this, I systematically compared domains of known structure using the state-of-the-art remote homology detection method HHsearch and identified 50 fragments that co-occur in domains with different folds, yet show significant similarities in sequence and structure. The occurrence of these homologous fragments in otherwise analogous structures provides compelling evidence for the antecedent domain segment theory. As an example, one of these 50 fragments, corresponding to a helix-strand-helix motif that gave rise divergently to three different folds, including the histone fold, is presented.

In addition to showing that most domains of one fold arose from an ancestral form by divergence, this thesis reveals many incidences of homologies between superfamilies of different folds due to the discovery of shared ancestral peptides. However, current protein classifications consider folds to be analogous and do not contain a hierarchical level to capture such inter-fold relationships. To solve this problem, this work proposes a classification level above the fold level, the metafold, which unites groups of folds for which a homologous relationship has been corroborated. The metafold level is an important step on the way to a classification of proteins by natural descent, which is the most informative basis for structural and functional inference.

Zusammenfassung

Obwohl die Vielfalt an Proteinen in der Natur grenzenlos zu sein scheint, ist sie in Wirklichkeit stark eingeschränkt. Viele Proteine haben eine erkennbare Ähnlichkeit in ihrer Sequenz und Struktur, da sie durch Amplifizierung, Rekombination und Divergenz aus einer Grundmenge sich autonom faltender Module, den Domänen, entstanden sind. Viele dieser Domänen gehen auf den letzten gemeinsamen Vorfahren (*engl. Last Universal Common Ancestor, LUCA*) zurück. Tatsächlich zeigt der Sequenzvergleich heutiger Proteine, dass man sie auf nur etwa 10000 Domänenfamilien zurückführen kann, die wiederum in nur etwa 3000 allgemeinere evolutionäre Superfamilien eingeteilt werden können. Darüber hinaus werden Superfamilien je nach topologischer Anordnung der Sekundärstrukturelemente einer von ungefähr 1000 Faltungen zugeordnet. Man geht davon aus, dass Faltungen analog sind, wobei Ähnlichkeiten zwischen verschiedenen Superfamilien einer Faltung das Resultat konvergenter Evolution sind. Allerdings haben das jüngste Anwachsen molekularer Datenbanken und Fortschritte in Sequenzvergleichsmethoden dazu geführt, dass viele entfernte evolutionäre Verwandtschaften, die über die Grenzen von Superfamilien hinausgehen, entdeckt wurden, was zeigt, dass nicht alle Superfamilien unabhängig voneinander entstanden sind. Das erste Ziel dieser Arbeit war es zu bestimmen, wie verbreitet solche entfernten Verwandtschaften sind. Dazu berechnete ich Cluster aus Domänen bekannter Strukturen entsprechend ihrer Sequenzähnlichkeit, des zentralen Kriteriums für die Ableitung eines gemeinsamen evolutionären Ursprungs. Die so entstandende Karte mit einer Vielzahl von Clustern zeigt, dass einerseits einige Faltungen mit vielen Superfamilien tatsächlich konvergent evolviert zu sein scheinen, und dass andererseits die meisten Domänen, die der gleichen Faltung angehören, aus einem Urprototyp entstanden sind. Dies zeigt, dass Proteine deutlich weniger polyphyletisch sind als bislang angenommen.

Die Entstehung heutiger Proteine durch eine kombinatorische Durchmischung einer begrenzten Anzahl an Domänen ist allgemein anerkannt; der Ursprung dieser Domänen ist jedoch nicht ausreichend verstanden. Selbst die simpelsten Domänen sind zu komplex, um *de novo* entstanden zu sein. Wenn dem so ist, wie sind dann die ersten Domänen entstanden? Diese Frage ist die Grundlage für das zweite Ziel meiner Arbeit. Eine Theorie über

die Entstehung der Domänen, die Theorie der ursprünglichen Domänensegmente, geht davon aus, dass Domänen durch Verschmelzung und Rekombination aus einer noch kleineren Auswahl an Peptiden enstanden sind, die als Kofaktoren aus der RNA Welt hervorgegangen waren. Dieser Theorie zufolge gingen aus dieser Auswahl durch Amplifizierung und Verschmelzung zunehmend stabilere Domänen hervor. Falls dies den Tatsachen entspricht, könnten viele moderne Proteine Überreste der Urpeptide, aus denen sie entstanden sind, beherbergen. Um dies zu erforschen, habe ich systematisch Domänen bekannter Struktur mithilfe der besonders empfindlichen Homologieerkennungsmethode HHsearch verglichen und 50 Fragmente identifiziert, die Domänen unterschiedlicher Faltungen zugeordnet sind, obwohl sie signifikante Ähnlichkeiten sowohl in ihrer Sequenz als auch in ihrer Struktur offenbaren. Das Auftreten dieser homologen Fragmente in ansonsten nicht homologen Strukturen stellt signifikante Evidenz für die Theorie der ursprünglichen Domänensegmente dar. Daraus folgern wir, dass diese Fragmente Überbleibsel der Urpeptide sind, aus denen die ersten Proteine hervorgingen. Als Beispiel wird hier eines dieser 50 Fragmente beschrieben, das einem Helix-Strang-Helix Motiv entspricht und in den Histonen sowie zwei weiteren Faltungen vertreten ist.

Diese Arbeit zeigt, dass die meisten Domänen einer Faltung durch Divergenz aus einer Urform entstanden, und identifiziert viele Homologien zwischen Superfamilien verschiedener Faltungen durch die Entdeckung gemeinsamer Urpeptide. Derzeitige Proteinklassifikationen gehen davon aus, dass Faltungen analog entstanden sind und sehen daher keine hierarchische Ebene vor, um solche Beziehungen zwischen Faltungen zu erfassen. Um dieses Problem zu beheben, führt diese Arbeit eine Klassifikationsebene oberhalb der Faltung ein, die Metafaltung, in der topologisch ähnliche Faltungen vereint werden, für die eine homologe Beziehung etabliert wurde. Die Metafaltung ist ein wichtiger Schritt auf dem Weg zu einer Proteinklassifikation gemäß natürlicher Abstammung, welche die informativste Grundlage für strukturelle und funktionale Inferenz ist.

Acknowledgements

Above all, I owe my deepest gratitude to my principal supervisor, Prof. Dr. Andrei N. Lupas, for providing the opportunity to work under his guidance, for sharing his unsurpassed knowledge on the evolution of proteins, and for his endless patience and kindness at all times. I have always been inspired by his extensive knowledge on a wide range of topics, including the history of science, music, and linguistics, his commitment to the highest standards of ethics, integrity, and quality, his lucid writing style, and his superlative presentation skills. Andrei's academic guidance has been invaluable over the years and I feel extremely privileged to have worked under his tutelage.

Next, I would like to extend my gratitude to my second supervisor, Prof. Dr. Oliver Kohlbacher, for co-supervising this thesis. His assistance has been priceless on an academic as well as a personal level. He was always easily approachable and available for discussions.

I am also highly indebted to Dr. Johannes Soeding, who supervised my work at the start of my doctoral studies. I would, in particular, like to thank him for developing the remote homology detection method HHsearch, without which this thesis would not have been possible.

This thesis has profited greatly from discussions and collaboration with a number of colleagues. I would, especially, like to thank Dr. Dirk Linke, Klaus O. Kopec, Marcus D. Hartmann, Martin Mechelke, Dr. Michael Habeck, Dr. Moritz Ammelburg, Dr. Murray Coles, and Dr. Stanislaw Dunin-Horkawicz. I would also like to thank all present and former members of the department of protein evolution for providing a stimulating and conducive research environment. I also owe a huge debt of gratitude to Ms. Karin Lehmann for her kindness, help, and support over the years.

My time in Tübingen was made enjoyable due to the love, affection, and support of many friends and I would, in particular, like to thank Ali Danish Zaidi, Anjan Hegde, Anirudh Ranganatha, Anurag Mishra, Arianne Chaudhuri, Arshiya Mishra, Avijit Pramanik, Beth Rowan, Bianca Priester, Carolin Ewers, Cedric Hobel, Christina Wassermann, Dara Forouzan, George Wang, Indronil Chaudhuri, Ivan Kalev, Jose Arcadio Farias Rico, Kalathur Ravi Kiran Reddy, Karsten Bogwardt, Markus Gruber, Martin Schueckel, Nagarajan Paramasivam, Nayantara Ranganatha, Ritu Mishra, Roosa Laitinen, Sitaram Ranganatha, Srinivasa Marimganti, Suryadeep Dash, Sus-

mita Pramanik, Vani Ranganatha, Vasuki Ranjani Chellamuthu, Vishal Kapoor, Vishnudev Ramachandra, Thomas Arnold, Thomas Holder, and Tobias Dezulian.

Lastly, I would like to thank my parents and my sister for their never-fading support, belief, and patience. I am deeply saddened by the fact that my father could not see this process come to end, but he is in my memories everyday.

In accordance with the standard scientific protocol, I will use the personal pronoun *we* to indicate the reader and the writer or (as mentioned in Appendix B) my scientific collaborators and myself.

Contents

Chapter 1

Introduction

1.1 Proteins - the bricks of life

Life in its present form is only possible because of the intricate and balanced interplay between the four major biological macromolecules: carbohydrates, lipids, proteins, and nucleic acids. Proteins are the most abundant of these macromolecules and account for more than 50% of the dry weight of living cells [Voet and Voet, 2004]. Given this abundance, it is not surprising that proteins carry out a multitude of vital functions in every living organism, be it bacteria or humans. In our bodies, they are central to almost all processes: they help us digest the food we eat, they form our hair and nails, they are important components of our bones, muscles, and tendons, they defend us from foreign invaders such as bacteria and viruses - in essence, we are what we are because of them.

Living organisms possess many widely diverse proteins that range in numbers from a couple of thousands in bacteria to many thousands in vertebrates. With about 10^8 species on earth and about 10^4 protein-coding genes per species [Bull et al., 1992], the estimated total complement of the world's proteome is approximately a trillion. In addition to being numerous, proteins also come in a wide array of shapes and sizes; for example, while hemoglobin, the protein that carries oxygen in our blood, is compact and round, collagen, the protein that forms a major component of our connective tissues, is long and rope-like (Fig. 1.1).

Despite their large numbers, various forms, and diverse functions, all proteins found in nature are made from the same ubiquitous set of 20 building blocks, known as amino acids. Proteins typically contain various combinations of 50 to 3000 amino acids connected linearly through a covalent peptide bond and thus they are frequently also referred to as polypeptides. All amino acids have a common basic structure comprising a central carbon atom that is covalently connected to a hydrogen atom, a carboxyl group, and an amino group, but differ due to the presence of a side-chain, which varies

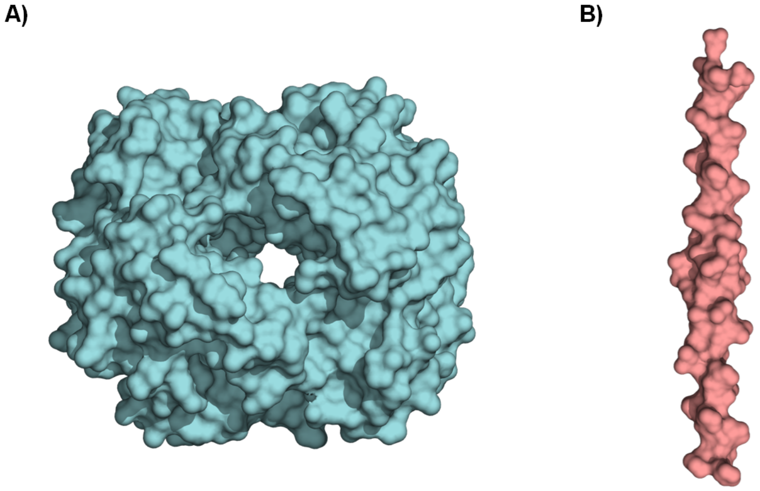


Figure 1.1: The diverse shapes and sizes of proteins. A) Hemoglobin is an oxygen-transport protein found in the red blood cells of most vertebrates, including humans. It is round and compact in shape, and is composed of four chains, two α and two β. In humans, the α chain and β chain comprise 141 and 146 amino acid residues, respectively. The structure shown is of human hemoglobin (PDB 1GZX). B) Collagen, the main component of connective tissues, is the most abundant protein in mammals. For instance, about one quarter of all proteins in humans is collagen. It is composed of three chains that are typically over 1200 amino acids long and are wound together to form a rope-like structure. The illustration shows a synthetic peptide containing a short region from human type II collagen (PDB 1BKV).

from a single hydrogen atom in the amino acid glycine to two aromatic rings in tryptophan. It is the nature of the side-chain that makes each of the 20 amino acids unique and provides them with individual characteristics; for example, depending on the nature of the side-chain an amino acid can be water-loving or water-repelling.

1.2 Protein folding

Each protein has a unique order of amino acids, termed its primary structure, that is specified by the nucleotide sequence of the gene encoding it. Genes are discrete units of DNA, which is the hereditary material of most life forms and contains all of the information required to build and maintain an organism. In genes, amino acids are encoded by groups of three nucleotides, called codons. This correlation between the sequence of nucleotides in DNA and the sequence of amino acids in proteins is referred to as the genetic code. Since DNA is built up from the four nucleotides (A, C, G, and T), the total number of possible codons is 64; however, because proteins are built from a basic set of only 20 amino acids, there is some redundancy in the genetic code. Indeed, for most amino acids, there is more than one codon.

To synthesize proteins, cells follow a two-step procedure that first transcribes genes encoded in DNA into messenger RNA (mRNA) and then translates the mRNA into linear chains of amino acids. However, the newly synthesized linear chain is not biologically active as it is. To become functional, it must fold into a well-defined three-dimensional structure [Branden and Tooze, 1999]. This process is called protein folding and is driven by a hydrophobic collapse of the polypeptide chain into a minimal energy configuration, the native conformation, in which amino acids are arranged in a few local and recurring hydrogen-bonded elements (e.g. α-helices and β-sheets) [Pauling et al., 1951; Levinthal, 1968; Pace et al., 1996]. Most proteins get to their native structure largely autonomously, with some assistance from cellular folding factors, and all the information required for this is contained in their sequence [Anfinsen, 1973; Lee and Tsai, 2005]. Nonetheless, the three-dimensional structure of a protein can not be deduced just from its amino acid sequence; this problem is called the protein folding problem, and it represents one of the most important open challenges in biology.

Even though a number of fundamental rules governing the structures of protein have been gleaned out and, in many cases, coarse models of a protein's structure can be calculated, a general solution to the folding problem has remained elusive [Lupas, 2008]. Presently, the three-dimensional structures of proteins can only be determined using complex experimental methods such as X-ray crystallography and nuclear magnetic resonance spectroscopy. However, these approaches are slow and expensive. As a result, the rate at which structure databases are growing significantly lags behind the growth of sequence databases; presently the RefSeq protein sequence database contains 14,090,554 sequences [RefSeq, 2012], whereas the protein data bank (PDB) only contains 79,521 structures [PDB, 2012].

Sequence space is essentially infinite. Even just considering a protein of 100 residues, the number of all possible sequences (20^{100} or $\approx 10^{130}$) makes the estimated number of particles in the universe (10^{80}) seem infinitesimal. This raises an obvious question: do all of these conceivable sequences assume a folded structure? It would not appear to be so. Experiments to obtain folded proteins from random sequence libraries yielded a success rate of no more than one in a billion [Keefe and Szostak, 2001], suggesting that sequences encoding folded proteins in the space of all possible sequences are extremely rare. Indeed, often one or a small number of point mutations in natural proteins result in them being misfolded, frequently impairing their biological functions. The consequences of this can be devastating, as is evident from the fact that a number of well-known human degenerative diseases, such as cystic fibrosis, Alzheimer's, Parkinson's, and Huntington's diseases, are caused by protein misfolding [Reynaud, 2010]. If arriving at folded proteins is so difficult, how did nature construct the millions of folded proteins that together sustain life on our planet?

1.3 Origin of folded proteins

Nature got around the protein folding problem by following a modular approach. Instead of evolving new proteins from scratch each time, it evolved them by extensively reusing an already available basic set of folded modules, termed domains. These existing domains were amplified, shuffled, and modified, while preserving their basic fold, to make up the proteomes of organisms [Soeding and Lupas, 2003]. Over the course of evolution, proteins that arose from a shared ancestor altered their sequence by point mutations, insertions, and deletions of residues, giving rise to a series of evolutionarily related variants or homologs. Indeed, sequence comparison of contemporary proteins shows that they can be grouped into about 10,000 domain families [Hunter et al., 2009; Punta et al., 2012], with members of each family exhibiting similar sequences, three-dimensional structures, and functions, reflecting their descent from an ancestral prototype. Many of these domains are widespread in each of the three kingdoms of life, indicating that they arose from a common prototype that was already present at the time of the Last Universal Common Ancestor (LUCA), some 3.5 billion years ago. In many cases, domains derived from a common ancestor can diverge to a point where they only show residual sequence similarity but nonetheless have a shared structure and similar functional properties. Such distantly related domain families are further classified into broader evolutionary superfamilies; the 10,000 domain families observed in modern proteins form about 3000 superfamilies [Hunter et al., 2009; Punta et al., 2012].

The diversity exhibited by the sequences of proteins does not correspond to a similar diversity in structural forms (folds). Many domains exhibit similar folds, even in cases when they share no features indicative of a common ancestry. Such structural similarities either represent an extreme case of divergent evolution or a case of evolutionarily unrelated domains having converged on the same fold. Because only a limited number of folded conformations are available to the polypeptide chain, owing to biophysical constraint, structural similarity of domains in the absence of detectable sequence similarity is generally thought to be the result of convergent evolution. This duality between divergent and convergent contributions to the evolution of proteins is captured by grouping superfamilies into folds based on the similarity of their structures; the aforementioned 3000 superfamiles fall into one of about 1000 folds [Andreeva et al., 2008; Cuff et al., 2011]. This grouping of superfamilies into folds has highlighted an intriguing fact: a small number of folds comprise many superfamilies, whereas the remainder corresponds to just one superfamily each [Ponting and Russell, 2002].

For instance, of the 1,195 folds recognized by the structural classification of proteins (SCOP) [Andreeva et al., 2008], only 136 comprise two or more superfamilies. The TIM $(\beta/\alpha)_8$-barrel fold is one such populous fold that is adopted by many totally unrelated enzymes with no detectable sequence similarity [Wierenga, 2001]; the SCOP database classifies proteins with this fold into 33 superfamilies. In the absence of detectable sequence similarity, the structural similarities between superfamilies of one fold were long thought to have originated independently, by convergent evolution. However, the recent growth of sequence and structure databases and the development of sensitive sequence comparison methods have brought forward an increasing number of instances of distant evolutionary relationships between superfamilies of one fold, suggesting that folds might not be as polyphyletic as previously assumed. For instance, most superfamilies of the TIM $(\beta/\alpha)_8$-barrel fold have now been shown to exhibit sequence similarity indicative of a common ancestry [Copley and Bork, 2000; Nagano et al., 2002; Soeding, 2005]. Such relationships have also been reported for a few other folds (e.g. β-propellers [Chaudhuri et al., 2008]), but their pervasiveness in the fold space is still unclear.

1.4 Objectives and structure of this thesis

In this thesis, we exploit the two aforementioned developments, namely the tremendous growth of molecular databases and the availability of sensitive sequence comparison methods, to answer questions concerning the polyphyly, origin, and classification of proteins.

The first objective of this thesis was to investigate the contributions of convergent versus divergent evolution to the origin of protein folds. In other words, we wanted to understand whether superfamilies of one fold are a result of divergent evolution from a common ancestor or a product of multiple independent origins. Even if superfamilies of one fold arose from an ancestral prototype, they might have diverged to a point where they only retain weak signals for common ancestry. We therefore used the state-of-the-art remote homology detection method, HHsearch [Soeding, 2005], which is based on the comparison of profile hidden Markov models (HMMs), to infer homology. Profile HMMs incorporate position-specific amino acid and gap preferences derived from a multiple sequence alignment comprising distantly related proteins and thus increase the sensitivity of sequence-based comparisons. To investigate the polyphyly of folds, we clustered representative domains of known structure by their sequence similarity, as evaluated by HHsearch, and generated a map with a high-level view of the evolutionary relationships in the fold space. We discuss our findings in the first part of Chapter 3.

It is widely recognized that modern proteins evolved by amplification, recombination, and divergence, from a limited set of domains, but the origin of this set itself is poorly understood. The detection of many instances of sequence- and structure-similar fragments within domains of different folds and the ever-increasing examples of domains composed of multiple copies of a homologous repeat, suggest that domains themselves are divisible and might have arisen from the assembly of simpler peptides. We favor the proposition that the first folded domains arose by fusion and recombination from an ancestral set of peptides, which emerged as cofactors in the context of RNA-dependent replication and catalysis in the RNA world [Fetrow and Godzik, 1998; Lupas et al., 2001; Soeding and Lupas, 2003]. From this perspective, the local similarities seen in modern proteins might represent the observable remnants of such ancestral peptides. We reasoned that if this hypothesis is true, systematic searches should allow us to detect many more such remnants in modern proteins. To this end, we systematically compared domains of known fold types using a sequence- and structure-based approach. We present our results in the second part of Chapter 3.

In Chapter 4, we describe a scenario for the origin of the histone fold based on the presence of an ancestral peptide that it common to it and two other folds. In Chapter 5, we discuss a widespread motif, the GD box, that is found in both homologous and analogous contexts.

The classification of proteins based on natural descent, akin the *Linnaean* system for organisms, is still an unsolved problem. Such a system is highly desirable for the comparative studies of proteins, as homology frequently offers a reliable source of structural and functional information. As described earlier, current systems classify proteins by combining homologous criteria (sequence similarity) at lower hierarchical levels with analogous criteria (structural similarity) at upper levels. By doing this, these systems make the implicit assumption that homologous domains always have the same fold; however, this is not the case. In recent years, several cases of distant evolutionary relationships between domains of different folds have been revealed, either due to the discovery of homologous fold change events (e.g. circular permutation) or due to the detection of homologous fragments. Because current systems consider folds to be analogous, they fail to account for these aspects of protein evolution. In Chapter 6, we discuss this issue in detail and propose a classification level, the metafold, to unite groups of folds for which a common origin has been substantiated.

Chapter 2

Background

This chapter provides a review of the basic theory underlying topics addressed in the following chapters. We start with an introduction to the four levels of protein structure and to the structural, functional and evolutionary units of proteins, domains. Subsequently, we discuss the classification of proteins and the mechanisms by which proteins change their fold. We end this chapter by focusing on a hypothesis for the evolution of folded proteins.

2.1 Levels of protein structure

Proteins are linear sequences made from 20 types of amino acids (also referred to as residues), each with different physicochemical properties. This linear sequence assumes a well-defined three-dimensional structure, which is unique to each protein and dictates its biological activity. Therefore, in order to understand the function of proteins, it is crucial to comprehend their structure. The structure of proteins is organized into four hierarchical levels: *primary*, *secondary*, *tertiary*, and *quaternary* [Branden and Tooze, 1999].

Primary structure refers to the linear sequence of amino acids in a protein (Fig. 2.1). The amino acids constituting a protein are connected end-to-end by covalent peptide bonds, which are formed during translation by the condensation of the carboxyl group of one amino acid with the amino group of the next. Proteins are therefore also referred to as polypeptides. With the exception of disulfide bonds that are formed between non-consecutive cysteine residues and help stabilize higher-order structures, all covalent bonds in proteins define the primary structure. In contrast, the higher-order structures of a protein, namely secondary, tertiary, and quaternary structures, are stabilized by non-covalent interactions.

Secondary structure refers to highly regular local substructures formed by linear segments of a polypeptide chain. Proteins exhibit two main types of secondary structures: the α-helix and the β-strand [Pauling et al., 1951].

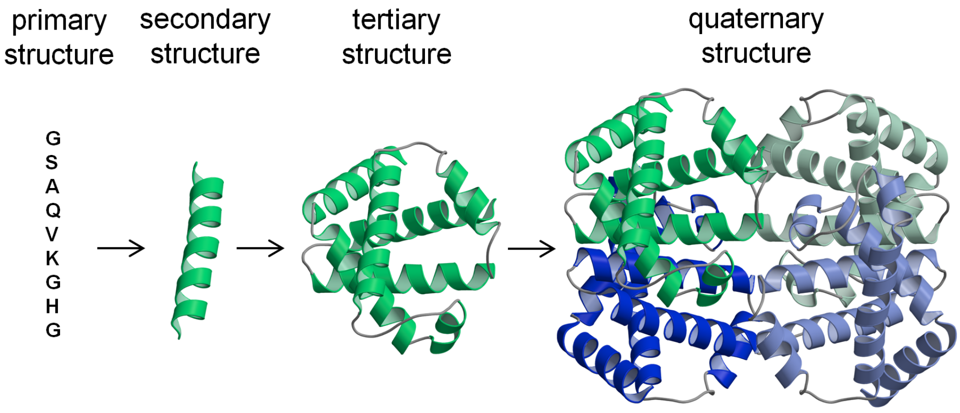


Figure 2.1: The four levels of protein structure. The structure shown is of human hemoglobin (PDB 2HHB), which is functional as a quaternary structure of four subunits, two α and two β . The α subunits are shown in shades of green and the β subunits in shades of blue.

The α-helix has 3.6 residues per turn arranged in a spring-shaped conformation, in which the CO group of each residue is hydrogen bonded to the backbone NH group of the fourth residue along the chain. In contrast, the β-strand is a fully extended segment of the polypeptide chain and is typically five to ten residues long. This extended conformation is unstable on its own and therefore single β-strands are rarely found in proteins. They are, more often than not, part of a stable arrangement called β-sheet, in which two or more β-strands are arranged next to each other such that hydrogen bonds can form between the CO and NH groups of adjacent β-strands. Most proteins are composed of combinations of α-helices and β-strands that are connected by loop regions of various lengths and shapes. Although less ordered than α-helices and β-strands, loops observed in proteins of known three-dimensional structure still fall into a limited set of geometries. Turns, a subtype of loops, assume somewhat ordered structures and are recognized as a third class of secondary structures. They are composed of three or four residues arranged in a compact, U-shaped element, which is stabilized by a hydrogen bond between residues at the first and last positions. In an average protein, 90% of the residues occur in α-helices, β-strands, and loops; segments that are not in these three secondary structures are referred to as random coils.

Tertiary structure refers to the overall folded conformation of a polypeptide chain (Fig. 2.1). The secondary structural elements, i.e. α-helices, β-strands, and loops, typically pack against each other in a compact fashion to give proteins their well-defined three-dimensional shapes. The main driving force behind the formation of the tertiary structure is the hydrophobic effect; globular proteins fold in a way such that they bury the side-chains

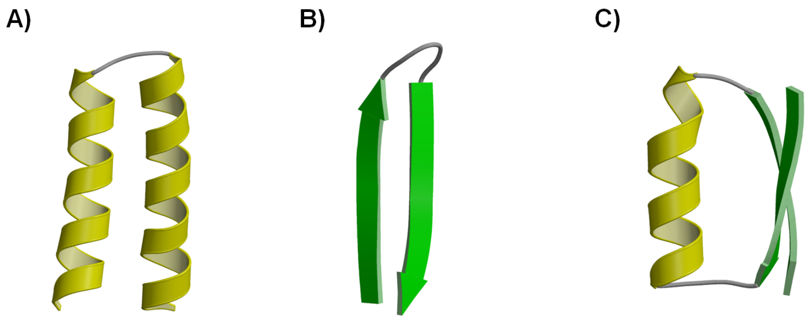


Figure 2.2: The three most recurrent supersecondary structural elements. In all elements α-helices are shown in yellow, β-strands in green, and loops in gray. The motifs shown are: A) a α-hairpin from the yeast vesicular transport protein Sec17 (PDB 1QQE), B) a β-hairpin from the snake venom erabutoxin (PDB 1QKE), and C) a β-α-β element from the NAD-dependent formate dehydrogenase protein from *Pseudomonas* sp. 101 (PDB 2NAC).

of nonpolar (hydrophobic) amino acids away from solvents in their core and expose the side-chains of polar (hydrophilic) amino acids to solvents. The situation is, however, different for membrane proteins since they reside in a hydrophobic environment, that is the lipid bilayer, instead of an aqueous one; they therefore expose hydrophobic residues to the lipid bilayer. In addition to hydrophobic interactions between the non-polar side chains buried in the interior, specific non-covalent interactions such as salt bridges and hydrogen bonds, and occasionally disulfide bonds between non-consecutive cysteine residues, also contribute to the stability of the tertiary structure of globular proteins. While for many proteins tertiary structure is the highest level of organization, some are only functional as higher-order assemblies.

Quaternary structure refers to complexes of two or more polypeptide chains (subunits) held together by non-covalent bonds. Complexes with identical and different subunits are called homo- and hetero-oligomers, respectively; e.g. hemoglobin is a hetero-tetramer of two α and two β subunits (Fig. 2.1).

Between the secondary and the tertiary level, protein structures exhibit a sublevel of organization called the *supersecondary structure* (also referred to as motif), which comprises two or more secondary structure elements in specific geometric arrangements. Certain supersecondary structures are widespread in proteins [Salem et al., 1999], e.g. α-hairpins (two adjacent α-helices connected by a short loop; Fig. 2.2A), β-hairpins (two adjacent β-strands connected by a short loop; Fig. 2.2B), and β-α-β motifs (two adjacent parallel β-strands connected by an α-helix; Fig. 2.2C). Such motifs frequently occur in evolutionarily unrelated proteins, indicating that their popularity is due to some general principles of the structural organization of proteins rather than due to common ancestry.

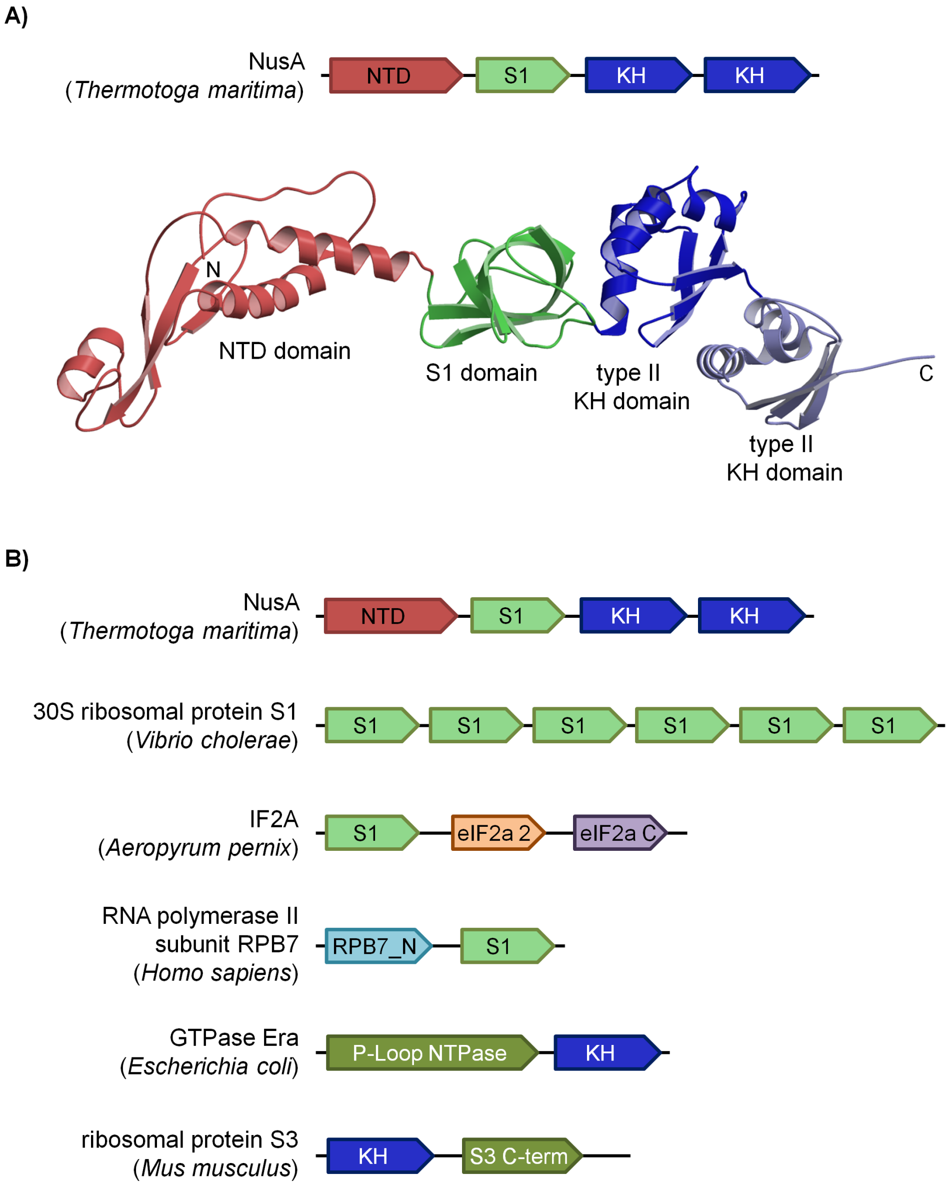


Figure 2.3: The modular nature of proteins. Modern protein are built of a basic set of recurring modules called domains. A) Domain architecture of the transcription factor NusA. The structure shown is of NusA from *Thermotoga Maritima* (PDB 1HH2). NusA is composed of four domains: an N-terminal α/β domain (shown in pale red), a five-stranded β-barrel (S1) domain (green), and two K-homology (KH) domains (blue). B) Domain architecture of proteins comprising the S1 and KH domains. Currently, the N-terminal domain of NusA is unique to it. The S1 and KH domains, however, reoccur in several proteins in different combinations with other domains.

2.2 Domains - the evolutionary units of proteins

Tertiary structures of large proteins are frequently composed of multiple physically distinguishable structural units, referred to as domains. Typical domains consist of 100 to 200 residues [Wheelan et al., 2000] in various combinations of α-helices, β-strands, and loops. These elements pack against each other in a compact fashion by burying their hydrophobic segments away from solvents in the core and by exposing their hydrophilic segments to solvents. This gives each domain its distinctive three-dimensional structure, termed fold [Cordes et al., 1996].

The individual domains of a multidomain protein typically fold more or less independently of each other. Furthermore, more often than not, isolated domains of a multidomain protein also fold into same tertiary structures as in the full protein. Therefore, besides being units of structure, domains are also units of folding. This property has been exploited extensively for determining the structures of large multidomain proteins, which are hard to crystallize, a prerequisite for X-ray crystallography, and are too big for nuclear magnetic resonance (NMR) techniques. To a large extent, such proteins have therefore been structurally characterized by determining the structure of their constituent domains, which are much smaller in comparison and are thus suitable for X-ray and NMR studies.

The smallest proteins are composed of a single domain, whereas larger proteins can contain as many as several dozen domains. Each subunit of hemoglobin, for instance, consists of a single domain, in which eight α-helices are packed against each other in an arrangement commonly referred to as the globin fold (Fig. 2.1) [Perutz et al., 1960]. Example of a multidomain protein includes the transcription factor NusA from *Thermotoga maritima*, which has an elongated rod-like tertiary structure composed of four domains: an N-terminal α/β domain, a five-stranded β-barrel (S1) domain, and two type II K-homology (KH) domains (Fig. 2.3A) [Worbs et al., 2001].

Typically, neither is a given domain found only once in a protein nor is it unique to it, instead it can reoccur multiple times in the same protein or in a variety of different proteins. Often the constituent domains of a protein carry out distinct functions and when they reoccur in other proteins, they frequently retain their function, and therefore in addition to being units of structure and folding, domains are also units of function. The previously described S1 domain, for example, is found in a large number of RNA-associated proteins, in which it is involved in interactions with RNA (Fig. 2.3B). Likewise, the K-homology (KH) domain, which occurs twice in NusA, is a domain of around 70 amino acids and is present in a wide range of nucleic acid-binding proteins; for example, in the eukaryotic ribosomal protein S3 (Fig. 2.3B). In contrast to the S1 and KH domains, the N-terminal domain of NusA is unique to it, as it has not yet been seen in any other protein.

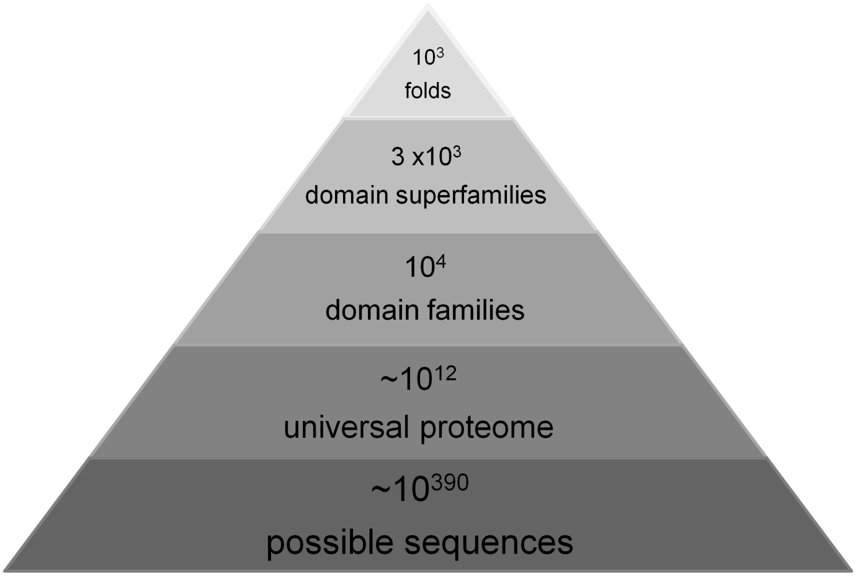


Figure 2.4: The protein universe. The sequence space of proteins is practically infinite. At a median protein chain length of about 300 residues, the number of conceivable protein sequences is 20^{300} (or $\approx 10^{390}$). With about 10^8 species on earth and 10^4 protein-coding genes per species, the total complement of the world's proteome is only about a trillion (10^{12}). As small as this number may seem by comparison to the combinatorial possibilities, it still represents a large overestimate of the actual protein diversity. Current protein classifications recognize 10^4 domain families by sequence similarity, which form about 3000 superfamilies, which in turn assume about 1000 topologically distinct structural forms (folds).

The recurrence of certain domain types in a wide variety of different multidomain proteins illustrates that modern proteins are built from simpler modules. This raises an obvious question: how big is the space of proteins and their constituent domains? Protein sequence space is essentially infinite (Fig. 2.4). Even just considering the median protein chain length of about 300 residues, the number of possible sequences is 20^{300} (or roughly 10^{390}), which vastly exceeds the estimated number of particles in the known universe (10^{80}). It is unlikely that nature could have explored more than a minuscule proportion of this enormous space. Indeed, the total complement of the world's proteome is probably only about a trillion (with about 10^8 species, each containing roughly 10^4 protein-coding genes [Bull et al., 1992]). Moreover, the trillion proteins present today are not a random sample of the polypeptide space. In fact, most proteins resemble other proteins in sequence and structure because they are built by amplification, recombination, and divergence from an already available, basic set of autonomously folding units, domains. Around 10^4 domain families, forming about 3000 broader evolutionary superfamilies (Fig. 2.4), have been recognized by sequence comparison [Hunter et al., 2009]. These superfamilies reflect the descent of modern proteins from a limited number of ancestral forms, most of which were already established at the time of the last common ancestor, 3.5 billion years ago. Diversity is even more reduced at the structural level. Often evolutionarily unrelated superfamilies exhibit the same fold and as a result the number of superfamilies does not correspond to an equal number of folds. Only about 1000 folds are populated in nature [Andreeva et al., 2008; Cuff et al., 2011]. Furthermore, many folds frequently show recurrent local arrangements of secondary structures (supersecondary structures; Fig. 2.2) [Salem et al., 1999], such that the diversity at the subdomain level is even further reduced. However, although the sequence similarity of domains reflects homologous descent, structural similarity may often be analogous because only a limited number of folded conformations are available to the polypeptide chain, owing to biophysical constraints.

This limited diversity of protein sequences and structures is not anecdotal. In fact, despite the continuing exponential increase of sequence and structure databases, the vast majority of new sequences and structures can be assigned to existing families and folds, and only few new families and no new folds have been recognized in recent years. This suggests that we have already achieved a fairly comprehensive view of the protein universe.

2.3 Classification of proteins

2.3.1 Overview

Over the last two decades, several different schemes for protein structure classification have been developed, and are routinely used for inferring the structure, function, and evolution of proteins. Because modern proteins evolved by the combinatorial shuffling of a basic set of folded domains, most systems, including the widely recognized CATH (Class-Architecture-Topology-Homology) [Cuff et al., 2011] and SCOP (Structural Classification of Proteins) [Andreeva et al., 2008] databases, use domains as the units of classification. Both CATH and SCOP attempt to capture the evolutionary and structural relationships between proteins by ordering them into hierarchies of families, superfamilies, and folds. Although philosophically similar, these databases differ in the way they are generated: while SCOP is mostly based on manual assignments, CATH employs automated and manual processes. We note that, in addition to CATH and SCOP, there are a number of other popular structural classification systems, examples of which include the FSSP [Holm and Sander, 1993] and the MMDB databases [Madej et al., 2012]. However, since SCOP, in particular, has become a key resource in the analysis of similarities and dissimilarities between proteins, we use it as a point of reference in this thesis.

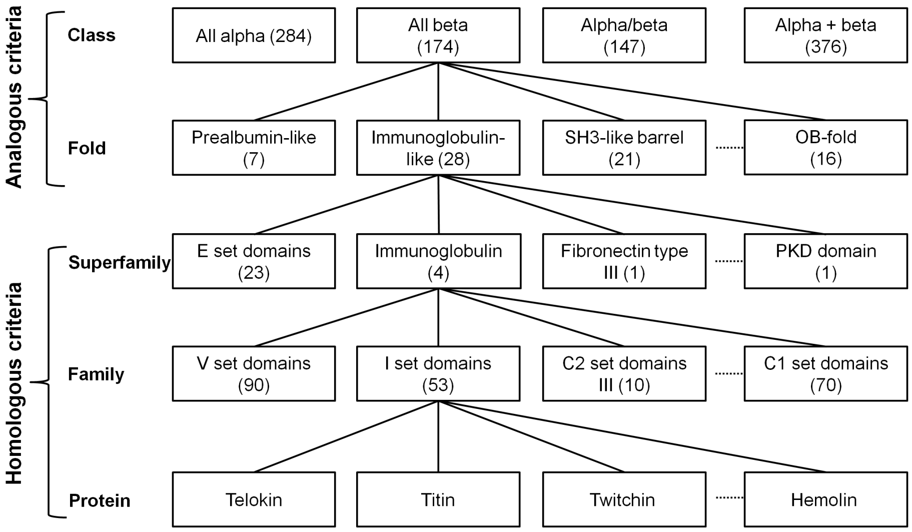


Figure 2.5: The hierarchical scheme of the SCOP database. Closely related domains are grouped together into families, which are in turn grouped into broader evolutionary superfamilies. Superfamilies are further assigned to folds based on the topological arrangement of secondary structure elements and folds are assigned to classes based on their secondary structure content. Hierarchical ordering of the immunoglobulin-like fold, which belongs to the all-β class and contains 28 superfamilies, is shown.

2.3.2 The SCOP database

The primary motivation of the SCOP database is to provide a comprehensive view of known structural and evolutionary relationships betweens proteins. This is achieved by organizing them into different hierarchical levels based on sequence and structural similarity. The fundamental unit of classification in SCOP is a domain. Proteins that comprise a single domain are regarded as one unit, whereas the domains of multidomain proteins are treated as separate units and are classified individually. The SCOP hierarchy comprises of six levels: species, protein, family, superfamily, fold, and class.

Species is the bottommost level of the SCOP hierarchy and corresponds to distinct domains and their experimentally generated variants. The next higher level, Protein, groups together orthologs from different organisms or isoforms from the same organism. Family comprises proteins that exhibit pairwise residue identities of 30% and greater. Occasionally proteins that have lower sequence identities, but extremely similar functions and structures are also placed together in a family. Superfamily encompasses families whose members have low sequence identities, but have structural and functional features indicative of a common ancestry.

The fold level separates superfamilies into groups that have the same overall tertiary structure (or fold). Domains are considered to have the same fold if their secondary structure elements are arranged and connected in the same way. Superfamilies of one fold have similar tertiary structures, but no sequence similarity indicative of a shared ancestry. It is thought that such superfamilies may have arisen by multiple independent events and that they adopt the same fold because of s reasons. However, it is also possible that superfamilies of one fold represent extreme cases of divergent evolution from a common ancestor. At the topmost level, folds are placed into four primary classes based on their secondary structure content: all-alpha (folds consisting primarily of α-helices), all-β (folds formed mainly of β-strands), α/β (folds in which helices and strands alternate regularly), and $\alpha+\beta$ (folds with irregular mixtures of helices and strands). SCOP also recognizes a small proteins class, which comprises proteins rich in cysteine, a multidomain class, which contains proteins with multiple domains, for which no homologs are known presently, and a membrane protein class, but these do not constitute classes in an architectural sense. The current version (1.75) of the SCOP database contains 110800 domains classified into 3902 families, 1962 superfamilies, and 1195 folds.

The example shown in Fig. 2.5 illustrates the hierarchical ordering of domains with the immunoglobulin-like β-sandwich fold [Bork et al., 1994]. This fold is formed of two β-sheets, each containing antiparallel β-strands, that envelop a central hydrophobic core. As this fold consists primarily of β-strands, it is classified into the all-β class in SCOP. The immunoglobulin-like fold comprises 28 superfamilies, each with multiple families. The immunoglobulin superfamily, for instance, contains four families: V set domains, C1 set domains, C2 set domains, and I set domains. Each of these families encompasses groups of orthologous proteins; the I set domains family, for example, includes the protein Tintin that is important in the contraction of striated muscle tissues.

Because SCOP groups domains by combining structural considerations in the upper classification levels (fold and class) with homologous criteria in the lower levels (family and superfamily), it makes the implicit assumption that homologous domains always have the same fold (Fig. 2.5). However, as has increasingly become clear in recent years, it is not always the case. Homologous domains can change their folds over the course of evolution owing to events such as circular permutation [Grishin, 2001a]. We present a few examples in *Section* 2.5.

2.3.3 Properties of folds

The hierarchical ordering of protein structures has brought an intriguing fact to light - a small number of structural forms (or folds) are far more common than others. For instance, of the 1,195 folds contained in SCOP, only 136 comprise two or more superfamilies, whereas the rest corresponds to just one superfamily each. Furthermore, only about 10 of these folds are really large and represent nearly 40% of all the sequence families in the PDB; these populous folds are referred to as superfolds [Orengo et al., 1994]. The aforementioned immunoglobulin-like fold is one of the best known superfolds and comprises a two-layered sandwich of between seven and nine antiparallel β-strands arranged in two β-sheets [Bork et al., 1994]. Domains of this fold are, in particular, ubiquitous in proteins of the immune system and in many cell surface receptors. SCOP classifies domains with this fold into 28 different superfamilies. Another well known superfold is the TIM $(\beta/\alpha)_8$-barrel, which is a toroidal fold comprising a closed eight-stranded parallel β-barrel surrounded by eight α-helices. Many totally unrelated enzymes with no sequence similarities adopt this fold and accordingly, SCOP classifies them into 33 different superfamilies within one fold. Why are certain folds, such as the immunoglobulin-like fold and the TIM barrel fold, highly populated? Many of them have simple topologies that either have internal structural symmetry or are composed to over 70% of the three most common supersecondary structures, α-hairpins, β-hairpins, and β-α-β motifs. It is therefore possible that the kinetic folding of these folds is efficient compared to that of more complex folds and as a result they may have been favored in the evolution of proteins [Orengo et al., 1994]. Furthermore, two scenarios have been discussed for the origin of superfamilies of one fold [Orengo and Thornton, 2005]. One view maintains that superfolds support a wider range of possibilities in the sequence space and thus superfamilies with the same fold were arrived at independently multiple times in the evolution of proteins. A contrary view, however, regards that superfamilies of one fold represent extreme cases of divergent evolution from a common ancestor. In fact, many recent studies have found evidence in favor of the second view and have substantiated homologous relationships between superfamilies of some populous folds; examples include the superfamilies of the TIM barrel fold [Copley and Bork, 2000; Nagano et al., 2002; Soeding, 2005] and the β-propeller fold [Chaudhuri et al., 2008].

Another striking property of folds is their permanence over the course of evolution. Even the simplest organisms contain domains representative of most of the about 1000 folds found in nature. In fact, the comparative analysis of proteomes of organisms from diverse branches of life shows that the majority of folds observed today were already established at the time of the Last Universal Common Ancestor (LUCA), some 3.5 billion years ago, and that very few novel folds arose later in bacteria, archaea, and eurkaryotes [Caetano-Anolls et al., 2009]. Examples of folds that were established in the LUCA and are preserved till date include the folds observed in the core complement of ribosomal proteins, which are still 40% identical between all organisms, ranging from amoeba to man [Ban et al., 2000]. Another example of extreme conservation is presented by the four nucleosome core histone proteins (H2A, H2B, H3, and H4), each of which is highly conserved from basal eukaryotes to higher mammals. Moreover, all core histone subunits have a common fold (the histone fold), in spite of only sharing low sequence similarity; this suggests that all core histones arose from a common ancestor [Arents et al., 1991]. This extreme conservation of protein structures is attributed to the discrete nature of the fold space [Sadreyev et al., 2009]. While the space of all possible protein sequences is essentially infinite, the number of sequences with a folded structure in this astronomical space is extremely small. The discovery of new folds is thus a really rare event. As a result, nature tends to work with already existing folds types, instead of evolving new ones. It adapts them to new functional demands by point mutations, insertions and deletions, while preserving their structures over billions of years.

2.4 Homology detection

The homologous descent of proteins from hypothetical ancestral forms is generally inferred from the similarity of modern representatives. These comparisons are typically carried out using sequence data, because sequence space is essentially infinite and convergence by chance therefore highly unlikely. Thus, proteins that exhibit statistically significant sequence similarities are reasonably considered to be homologous. However, for distant evolutionary events, sequences may have diverged beyond our ability to detect their evolutionary relatedness. Because the three-dimensional structures of proteins diverge much more slowly compared to their sequences, structure similarity is often used as a measure to identify such distant events. However, similar structures may have arisen convergently from different origins and their similarity thus frequently does not provide conclusive evidence of common ancestry [Salem et al., 1999; Cheng et al., 2008]. Nevertheless, structure comparisons may provide important clues to deep evolutionary relationships and seem particularly useful when coupled to sequence comparisons.

In the last two decades, inference of homology from sequence comparisons has proven to be extremely reliable for extrapolating biological knowledge. As a result, sequence comparison methods have become one of the most important tools in molecular biology. Sequence comparison methods achieve varying levels of sensitivity based on the amount of information they include. Sequence-to-sequence comparison methods such as BLAST [Altschul et al., 1990] and FASTA [Pearson and Lipman, 1988] that compare single sequences, scored by a global substitution matrix, are the least sensitive. An increased sensitivity is achieved by profile-to-sequence comparison methods such as PSI-BLAST [Altschul et al., 1997], which represent the query sequence using a sequence profile derived from a multiple alignment of homologous sequences. Sequence profile includes more information about the sequence family than a single sequence. In particular, it records the probability for the occurrence of each amino acids at every position of the multiple alignment. Since position-specific amino acid preferences are conserved over larger evolutionary distances than amino acid identities, profile-to-sequence methods exhibit greater sensitivity over sequence-to-sequence methods. An additional improvement in sensitivity is obtained by profile-to-profile comparison methods, such as COMPASS [Sadreyev and Grishin, 2003], which use family-specific information as represented by profiles for both sequences being compared.

Inclusion of position-specific insertion and deletion probabilities into profiles yields profile hidden Markov models (HMMs) [Eddy, 1998]. Methods based on the comparison of profile HMMs are presently the most sensitive tool for the detection of homology. Like position-specific amino acid preferences, position-specific insertion and deletion probabilities are also conserved over larger evolutionary distances, and this is the reason for the increased sensitivity of HMM-to-HMM comparison methods over profile-to-profile methods. HHsearch [Soeding, 2005; Soeding et al., 2006], an HMM-to-HMM comparison method, has a sensitivity equivalent to that of advanced fold recognition methods, in spite of using only sequence information. In fact, HHsearch was among the best-performing servers in the most recent CASP[1] experiment (Critical Assessment of Techniques for Protein Structure Prediction). HHsearch is currently one of the most sensitive methods for remote homology detection and has been used in numerous studies, by us and others, for identifying evolutionary relationships between protein families previously thought to be unrelated [Coles et al., 2005; Alber et al., 2007; Gao et al., 2007; Remmert et al., 2009; Kopec et al., 2010]. We have therefore applied HHsearch in this thesis to evaluate evolutionary relationships between distantly related proteins.

HHsearch can be employed either to perform pairwise comparisons between two proteins of interest or to search a standard protein database, such as the protein data bank (PDB), for homologs, starting with a single protein. As noted in the previous paragraph, HHsearch represents both query and database proteins by profile HMMs. Before initiating the search process, HHsearch generates a multiple alignment for the query protein either with PSI-BLAST or with a context-specific version of PSI-BLAST, called CSI-BLAST [Biegert and Soeding, 2009]. Next, the resulting alignment is assigned predicted secondary structure using the program PSIPRED [Jones, 1999]. A profile HMM representing the query protein is then derived from this annotated alignment. Profile HMMs for proteins in the databases provided by HHsearch are pre-calculated in a similar way; however, secondary structure information contained in these profile HMMs are either predicted by PSIPRED or are calculated from three-dimensional structure

[1] http://predictioncenter.org/casp9/CD/data/html/groups.server.tbm.html

using DSSP [Kabsch and Sander, 1983]. This incorporation of secondary structure information in query and database profile HMMs improves the sensitivity of homology detection [Soeding, 2005]. The output of HHsearch contains a ranked list of database matches and the corresponding pairwise sequence alignments. Every database match is assigned a probability value, which is an estimate for the likeliness of the given match to be a true positive. Generally, at probability values of $> 50\%$ the error rate of HHsearch is very low, that is, the detected relationships are most certainly homologous.

2.5 Fold change in the evolution of proteins

The evolutionary stability of proteins folds is astonishing. In fact, the vast majority of popular folds seen in modern proteins were already established at the time of the Last Universal Common Ancestor [Caetano-Anolls et al., 2009] and have not changed much ever since. However, this does not imply that proteins never alter their structural forms. It is long known that over the course of evolution, domains derived from a common ancestor can change by embellishing their structures, that is, by acquiring additional segments at their N- and/or C-terminal ends. Typically, these changes are minimal and do not alter the overall fold of domains. However, occasionally homologous domains can also undergo drastic structural changes, wherein one or more secondary structure elements within the fold alter their topology, leading to a change in the overall fold. In recent years, several events have been described by which homologous proteins can change their fold substantially [Grishin, 2001a; Andreeva and Murzin, 2006; Andreeva et al., 2007; Alva et al., 2007, 2008]. For example, fold change can result from the cumulative effect of point mutations (Fig. 2.6D) or by the insertion and deletions of segments (Fig. 2.6B). Other mechanisms for fold change include topological substitutions, strand swaps, strand and hairpin invasions, circular permutations, and 3D domain swaps; most of these have been discussed in detail by *Grishin et al.* [2001a].

Circular permutation, for instance, is a process that proceeds by a fusion of the N and C termini and a cleavage at a different position to create new termini. This results in proteins with different topological connections, but with similar overall structures (Fig. 2.6A). Sometimes, the cumulative effect of mutations, insertions, and deletions can cause homologous proteins to change their structures dramatically, even to the extent where they show no resemblances in structure. Such a case of dramatic fold change is presented by the transcription factor NusG and its paralog RfaH [Belogurov et al., 2007]. NusG is composed of two independent domains, an N-terminal domain (α/β fold) that has been implicated to bind RNA polymerase (RNAP) and a C-terminal β-barrel domain. While the N-terminal domain of RfaH resembles the corresponding domain in NusG, its C-domain

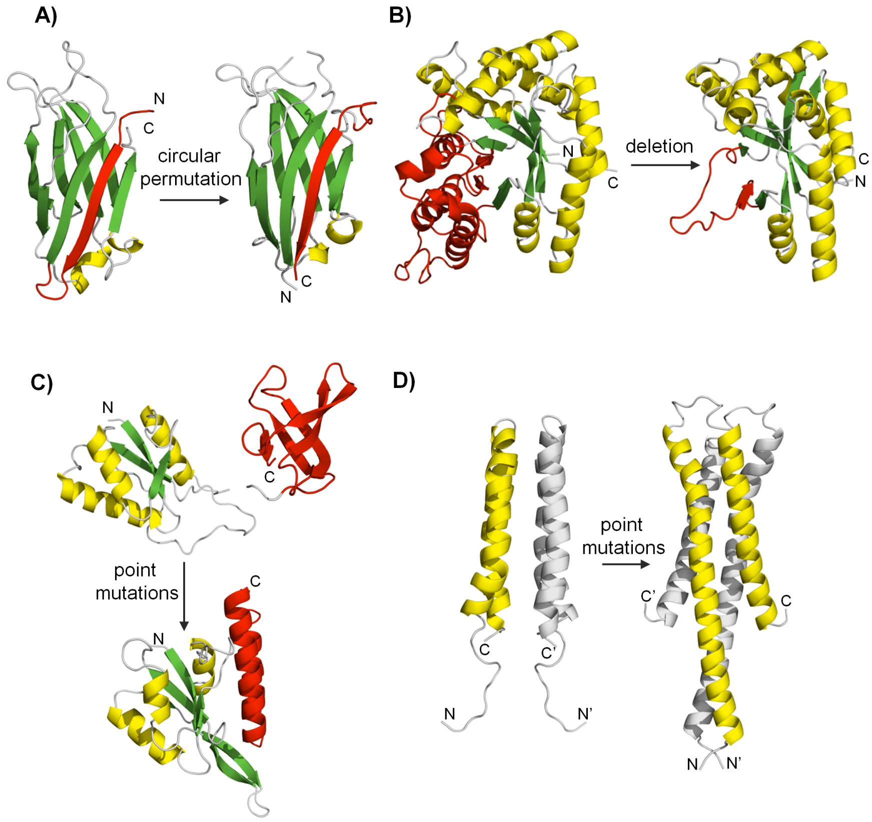


Figure 2.6: Mechanisms of fold change. In all structures α-helices are shown in yellow and β-strands in green; in homo-dimers, one monomer is shown in gray. A) The C2 domains of synaptogamin I (PDB 1RSY) and phospholipase C (1QAS) are related by a circular permutation event; the permuted strand is shown in red. B) The eight-stranded TIM barrel of bacterial luciferase (1LUC) is connected to the seven-stranded barrel of the nonfluorescent flavoprotein (1NFP) by deletion and helix-to-strand transition; the affected region is shown in red. C) The transcription factor NusG and its paralog RfaH. NusG contains two independent domains, a N-terminal domain and a C-terminal β-barrel domain. The N-terminal domain of RfaH resembles the N-domain of NusG, but its C-domain is a two helical coiled-coil; the affected region is shown in red. Although the C-domains assume different structures, they have about 20% sequence identity. D) The cumulated effect of point mutations resulted in a transition in the handedness of a four-helix bundle between the DHp domains of EnvZ (1JOY) and *Thermotoga* TM0853 (2C2A).

is a two helical coiled-coil (Fig. 2.6C). Despite the differences in the structure, the C-domains show about 20% sequence identity, indicating that they are homologous. Many residues that make up the hydrophobic core of the NusG β-barrel are also conserved in the RfaH coiled-coil and form a hydrophobic surface that interacts with the proposed RNAP-binding segment of the N-domain. This fold change event also resulted in a more evolved function for RfaG. NusG is a general transcription factor, whereas RfaH has a more regulated function; it only exposes its RNAP-binding site when certain sequence-specific interactions with DNA take place. This and other examples of fold change events illustrate the fact that nature frequently creates new structural forms by modifying existing ones, rather than building them from scratch.

2.6 Origin of domains

It has been long recognized that nature evolved contemporary proteins by shuffling existing domains types, which it customized to new functional demands by point mutations, insertions, and deletions. Many of these domain types were already established at the time of the LUCA, with a few domains arising later in bacteria, archaea, or eukaryotes [Rost, 2002; Caetano-Anolls et al., 2009]. Domains therefore represent units of evolution in modern proteins. While the origin of multidomain proteins has been researched and discussed extensively, the emergence of domains themselves is poorly studied.

Several lines of evidence, including the identification of many instances of local sequence and structure similarities within domains of different folds and the detection of numerous examples of folds composed of sequence- and structure-similar repeats, suggest that domains themselves are divisible and thus might not constitute the fundamental evolutionary atoms of proteins. In 2001, *Lupas et al.* [2001] examined these observations in detail and hypothesized that the first domains may have arisen by amplification and accretion from a small pool of simpler modules, much in the manner how modern multidomain proteins arose from a limited set of domains[2].

First, it was argued that the combinatorial complexity of creating a domain from scratch is unapproachable [Soeding and Lupas, 2003]. Even for a domain of just 100 residues, there are 20^{100} possible sequences. However, as demonstrated by experiments that tried to obtain folded proteins from random sequence libraries, only a minuscule fraction of these conceivable sequences will assume a folded structure [Keefe and Szostak, 2001]. This raises the question of whether nature could have sampled the astronomical space of all possible protein sequences in a reasonable amount of time to isolate folded domains. It does not seem so. Even if we assume that nature could have sampled the sequence space at an unrealistically fast rate of one trillion (10^{12}) different sequences per second, it would still take 4.02×10^{110} years to try all possible sequences to evolve a domain of 100 residues; to put this number in perspective, consider that only 13.7 billion (13.7×10^{9})

[2] For a detailed account on *the origin of domains*, please refer to [Lupas et al., 2001] and [Soeding and Lupas, 2003].

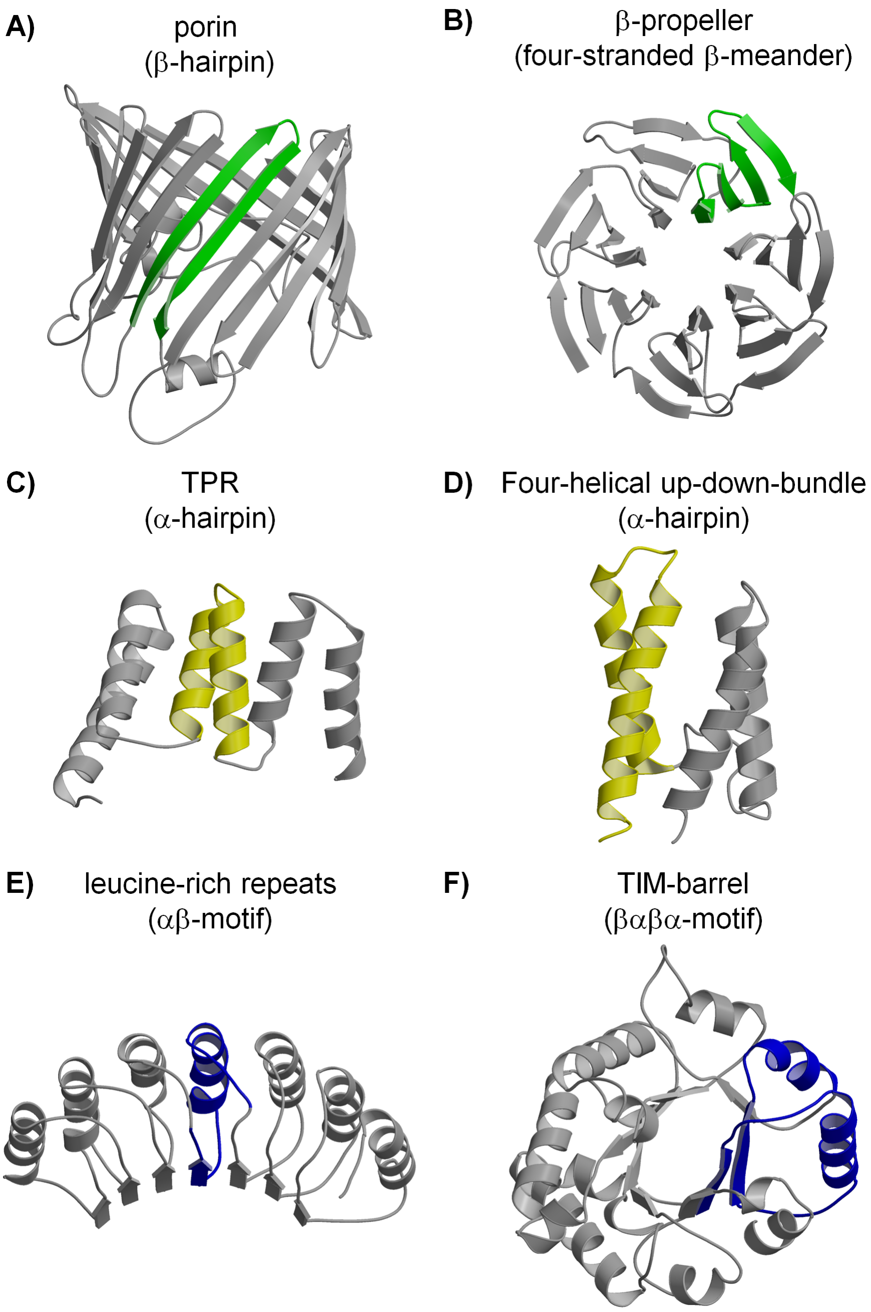
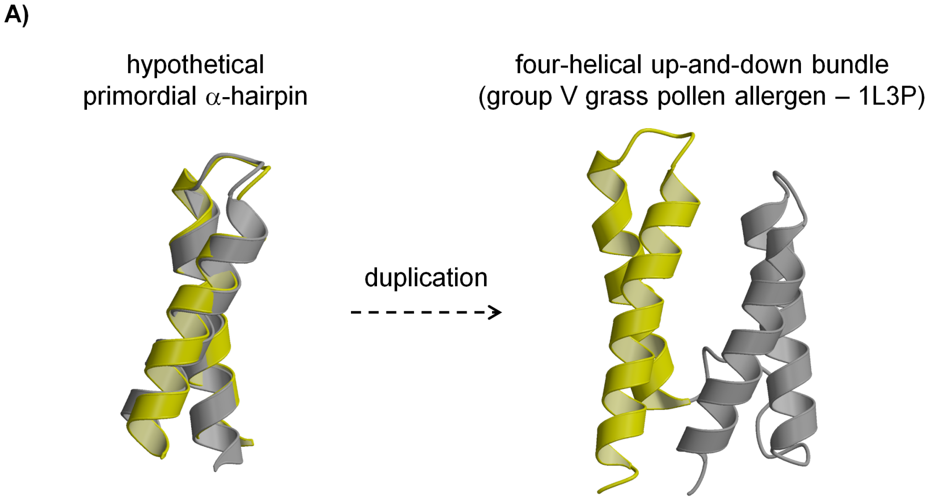


Figure 2.7: Folds with internal symmetry. α-hairpins are shown in yellow, β-hairpins in green, and motifs with mixed α and β in blue. The folds shown are: A) porin (PDB 2POR), B) β-propeller (1TBG), C) TPR (1ELR), D) four-helical up-down bundle (1L3P), E) leucine-rich repeats (2BNH A:118-318), and F) TIM $(\beta/\alpha)_8$-barrel (1HTI).


A)
hypothetical
primordial α-hairpin
duplication
four-helical up-and-down bundle
(group V grass pollen allergen – 1L3P)

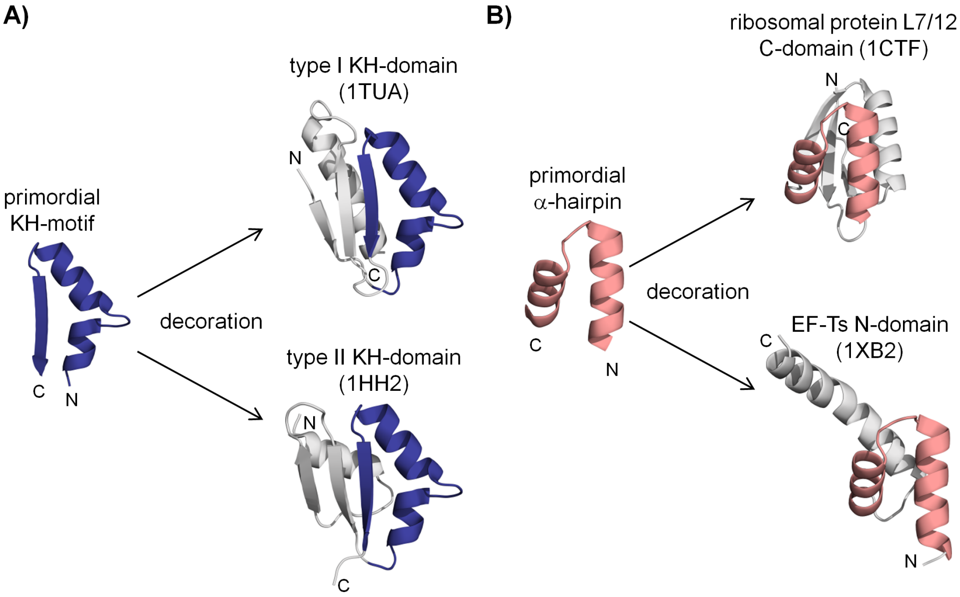
B)

```
                    hhhhhhhhhhhhhhhhhllllllhhhhhhhhhhhhhhh
1L3P (repeat 1)     IIDKIDAAFKVAATAAATAPADDKFTVFEAAFNKAIKE
1L3P (repeat 2)     CIPSLEAAVKQAYAATVAAAPQVKYAVFEAALTKAITA
```

Figure 2.8: Evolution by duplication. This example illustrates the evolution of a four-helical up-down bundle domain through duplication of a α-hairpin. A) Structural superposition of the two repeats of the four-helical up-down bundle (PDB 1L3P) is shown. The repeats are shown in yellow and gray. B) Sequence alignment of the two repeats. Conserved residue columns are shown in boldface and the secondary structure composition is indicated above the alignment.

years have passed since the Big Bang. It thus seems highly improbable that domains were evolved *de novo* from the gigantic space of all possible sequences. Furthermore, even if random sampling of the sequence space were possible, no non-biological processes are known that could synthesize and supply polypeptide chains of sufficient lengths and in quantities required for this exploration. However, abiotic synthesis of short peptides has been demonstrated [Keller et al., 1994]. It was, therefore, argued that the evolution of proteins started from short, rather than long peptides.

Second, it was reasoned that domains comprising multiple copies of a repeat may be the result of evolution by amplification and gene fusion from an ancestral homomultimer (Fig. 2.7). This idea had also been put forward by a number of previous studies, the earliest of which were carried out by Andrew McLachlan in the 1970s. He reported many examples of domains composed of repeats and proposed that repetition is an important mechanism in the evolution of multidomain proteins as well as in the evolution



C) **RNA-binding KH-motif**

PDB accession	SCOP ID	hhhhhhhhhhhhccchhhhhhhhhh
1XB2 (B:59-83)	a.5.2.2	SKE**LL**MK**LR**RK**TGY**S**F**IN**CK**K**ALET**
1CTF (A:65-89)	d.45.1.1	KVA**VI**KA**VR**GA**TGL**G**L**KE**AK**D**LVES**

D) **EF-Tu-binding α-hairpin**

PDB accession	SCOP ID	chhhhhhhhcchhhhhhhhhhhhceeeeee
1TUA (A:11-39)	d.51.1.1	**P**ER**L**G**AVIG**PR**G**EVKAE**IMR**R**TG**TV**I**T**V**D
1HH2 (P:311-339)	d.52.3.1	**P**TQ**L**S**LAIG**KG**G**QNARL**AAK**L**TG**WK**I**D**I**K

Figure 2.9: Sequence- and structure-similar fragments in domains of different folds. A) The type I KH domain (PDB 1TUA, A:3-72) and the type II KH domain (1HH2, P:277-342) share a homologous motif (shown in blue), which is involved in RNA-binding. These two domains may have arisen from an ancestral peptide through accretion. B) The C-domain of ribosomal protein L7/L12 (1CTF) and the N-domain of EF-Ts (1XB2, B:56-111) have a common α-hairpin (shown in pink). This motif is involved in interactions with EF-Tu. C) and D) show sequence alignments for the RNA-binding KH-motif and the EF-Tu-binding α-hairpin, respectively. Conserved alignment columns are shown in boldface and secondary structure composition is indicated above the alignment.

of domains themselves [McLachlan, 1979; McLachlan et al., 1980; McLachlan, 1980]. This proposition was primarily based on internal symmetries observed in folds at the structural level, a feature that generally reflects analogy. However, many studies have since presented numerous examples of folds that are built of repeats similar both in sequence and structure, confirming the role of repetition in the evolution of domains. One example for evolution by amplification is displayed by the group V grass pollen allergen which assumes a four-helical up-down bundle composed of two sequence- and structure-similar α-hairpin repeats (Fig. 2.8). This suggests that the ancestral form of this fold was a dimer of identical subunits, which fused to form a single chain version. It is also thought that stable assemblies of identical subunits evolve more quickly than those involving nonidentical subunits, owing to symmetry-related reasons [Lukatsky et al., 2007]. Indeed, of the ten most populated folds described by *Orengo et al.* [1994], six show internal structural symmetry. For instance, β-propellers, which are toroidal structures and are classified into five different folds comprising 32 superfamilies in SCOP, are built of between four and ten homologous repeats of a four-stranded β-meander (Fig. 2.7B) [Chaudhuri et al., 2008]. Examples of other highly populated symmetric folds include the TIM $(\beta/\alpha)_8$-barrel fold (Fig. 2.7F) [Copley and Bork, 2000; Nagano et al., 2002; Soeding, 2005] and the outer membrane β-barrel fold (Fig. 2.7A) [Remmert et al., 2009]. Some other highly populated folds, such as immunoglobulin and jelly-roll folds, show clear internal structural symmetry, but no corresponding sequence symmetry. At this time it is unclear if these structurally repetitive folds emerged by amplification form an ancestral module or whether their symmetric structure is an outcome of convergent evolution. This, until recently, was also the case for TIM barrels and β-propellers, as their repeats did not show detectable sequence similarity. However, the recent growth of molecular databases and the availability of sensitive sequence comparison methods have revealed that their repeats are in fact homologous. This might also turn out to be the case for other folds that presently only show internal structural symmetry.

Third, the detection of numerous instances of sequence- and structure-similar fragments within non-homologous folds provides further evidence for the notion that domains themselves arose from simpler modules. Examples of such locally similar substructures include the Asp-box [Copley et al., 2001], the DNA-binding helix-hairpin-helix-motif [Doherty et al., 1996], the KH-motif [Grishin, 2001b], and the EF-Tu-binding α-hairpin [Wieden et al., 2001]. The KH-motif (α-α-β), for instance, is found in two topologically distinct RNA-binding domains, the type I and the type II KH domains (Fig. 2.9A,C). The similarity between these two domains is limited to the KH-motif, which in both domains in involved in interactions with RNA. It is thought that the type I and type II KH domains arose from an ancestral module by accretion, that is, by multiple additions of structural elements. Another example of a sequence- and structure-similar motif is the α-hairpin common to the C- terminal domain of ribosomal protein L7/L12 and the N-domain of EF-Ts, both of which have otherwise different folds (Fig. 2.9B,D).

In both these domains, this motif is implicated to be involved in interactions with EF-Tu. The existence of these sequence- and structure-similar motifs in domains of different folds indicates that many domains might have originated by shuffling and accretion from simpler, ancestral modules.

Based on these observations, *Lupas et al.* [2001] postulated that the first folded domains might have arisen by fusion and recombination from an ancestral set of peptides [the antecedent domain segments (ADSs)] that emerged in the context of RNA-dependent replication and catalysis (the RNA world). According to this model, the local similarities found in modern proteins represent the observable remnants of such peptides. It was discussed that in the RNA world [Jeffares et al., 1998], which is widely accepted to be an important intermediate stage in the origin of the cellular life known to us today, simple peptides may have been recruited by RNA to expand its functional repertoire. While RNA was the central molecule in this pre-biotic life and carried out a wide range of functions, including storage of genetic information and catalysis of chemical reactions, there were some important reactions that could not have been performed by RNA on its own [Doudna and Cech, 2002; Joyce, 2002]. One such reaction is the redox reaction involving free radicals. In contrast to RNA, peptides are good chelators of small molecules and can assist redox reactions via their side chains. For this reason, peptides may have been chosen to assist RNA as cofactors in the pre-biotic world.

It is thought that initially short peptides were available by abiotic synthesis and that many non-specific RNA-peptide complexes may have formed spontaneously. This provided an avenue for the selection of RNA-peptide complexes with useful features, in terms of improved catalytic activity, higher thermostability, efficient folding, and ability of peptides to form secondary structure [Soeding and Lupas, 2003]. Over the course of evolution, many beneficial, specific RNA-peptide complexes may have been found and become integral to the RNA world. This in turn must have necessitated the need for a more straightforward synthesis of peptides and presumably led to the eventual emergence of a primitive genetic code and to the origin of single-gene minichromosomes [Jeffares et al., 1998]. The lengths of these mini-genes would initially have been restricted because of the high error rates of the early replication machinery. Consequently, the first protoproteins must have been either single, short peptides or oligomers of short peptides. Later, when the error rates of the replication process lowered, many of these mini-genes might have fused together, resulting in longer, single-chain genes, and thus proving the necessary raw material for the selection of first proteins that were capable of folding and functioning independent of their RNA template. These first proteins must have then diversified, essentially by addition of multiple structural elements (accretion), to give rise to the domain superfamilies and the fold types we recognize today.

Chapter 3

On the polyphyly and origin of domains

3.1 Motivation

All forms of life on earth, from the microscopic amoeba to the gigantic whale, are evolutionarily related to each other because they descended from a single ancestor, usually referred to as the Last Universal Common Ancestor (LUCA) [Glansdorff et al., 2008]. Life is therefore monophyletic and can be represented using a single phylogenetic tree with the LUCA at the root. The situation with proteins is, however, different. Modern proteins did not descend from a single ancestor, instead they evolved by amplification, shuffling, and divergence from a basic set of distinct ancestral prototypes (domains), many of which were already present at the time of the LUCA. Proteins are therefore polyphyletic and cannot be represented by a single tree, as for organisms. This polyphyly is evidenced by the fact that sequence comparisons of modern proteins recognize around 10,000 domain families, which further form about 3000 broader evolutionary superfamilies [Hunter et al., 2009; Punta et al., 2012], each encompassing descendants of an ancestral form. This variability in sequences does not correspond to a similar variability in structures. The three-dimensional structures of proteins exhibit regularities, and many domain superfamilies exhibit the same or similar folds, even in cases when they have no obvious evolutionary relationship. Such structural similarities may often be analogous because only a limited number of folded conformations are available to the polypeptide chain, owing to biophysical constraints. In line with this, the 3000 superfameles observed in modern proteins fall into one of about 1000 folds based on the similarity of their tertiary structures [Andreeva et al., 2008; Cuff et al., 2011], with the structural resemblances between superfamilies of one fold being the result of convergent evolution.

Parts of this chapter have been adapted with permission from [Alva et al., 2010].

How comprehensive is this polyphyly of proteins? Did each of the 3000 or so superfamilies arise independently, the structural similarities between them being convergent? This does not appear to be the case. In recent years, the dramatic growth of protein sequence and structure databases and the advances in sequence comparison methods have led to the discovery of many distant evolutionary relationships that transcend the boundaries of superfamilies, suggesting that not all protein superfamilies had arisen independently. The most noteworthy case is that of the TIM (β-α)$_8$-barrel fold, which is a toroidal fold comprising a closed eight-stranded parallel β-barrel surrounded by eight α-helices. This fold is the most popular fold among enzymes [Wierenga, 2001] and is seen in many different enzyme families, each catalyzing completely unrelated reactions. Although these families exhibit the same fold, the sequence similarity between them is weak, and they were therefore thought to have converged on the TIM-barrel fold independently. The SCOP database, for instance, classifies families with this fold into 33 analogous superfamilies. However, with the recent availability of sensitive sequence comparison methods, superfamilies of the TIM-barrel fold have been shown to exhibit sequence similarity indicative of homology, leading to the notion that they arose from a common ancestor [Copley and Bork, 2000; Nagano et al., 2002; Soeding, 2005]. This is also the case for a number of other highly populated folds, including β-propellers [Chaudhuri et al., 2008] and outer membrane β-barrels [Remmert et al., 2009]. In addition to the detection of instances where superfamilies of the same fold are homologous, occasionally, even the boundaries between folds have been broken either due to the discovery of homologous fold change [Grishin, 2001a; Andreeva and Murzin, 2006; Andreeva et al., 2008; Alva et al., 2008] owing to events such as circular permutation and strand invasions or due to the detection of conserved supersecondary structures, which may represent remnants of an ancient peptide-RNA world [Fetrow and Godzik, 1998; Lupas et al., 2001; Soeding and Lupas, 2003]. These findings suggest that proteins might not be as polyphyletic as hitherto assumed.

To evaluate the extent to which such distant relationships transcend current structural classification, we clustered a representative set of protein domains, encompassing all known folds, on the basis of sequence comparisons alone. The resulting map shows that many protein families from different superfamilies (see *Section* 3.2) or even folds may have a homologous origin (see *Section* 3.3).

3.2 An homology-based clustermap of folds

3.2.1 Previous studies

The nature of the 'fold space' has been a matter of intense research for decades and hitherto, a number of studies have mapped it by structural criteria, mainly with a focus on principles for automatically classifying proteins [Holm and Sander, 1993; Orengo et al., 1993; Hou et al., 2003]. *Holm et al.*, for instance, performed an all-against-all comparison of about 200 representative structures with less than 30% pairwise sequence identity, using their then newly-developed structure alignment algorithm DALI, and showed that an objective classification of proteins into structural classes (such all-α and all-β), as defined by visual inspection, is achievable [Holm and Sander, 1993]. Some studies also considered structure-based function inference [Hou et al., 2005] and the fold usage of organisms [Hou et al., 2003]. *Kim and coworkers* clustered protein structures based on their pairwise structural similarities to generate a high-level map of the fold space and showed that proteins with similar functions come to lie close together in this map [Hou et al., 2005]. All these studies used structural similarity to connect different folds, a property that primarily reflects analogy. As described earlier, unrelated domains frequently show recurrent local substructures (such as α- and β-hairpins), and because of these convergent local similarities, structural maps show proteins in a continuum [Friedberg and Godzik, 2005; Kolodny et al., 2006; Taylor, 2007; Cuff et al., 2009; Pascual-Garca et al., 2009], obscuring discrete evolutionary relationships [Sadreyev et al., 2009]. Indeed, recent results suggest that events such as circular permutations, strand invasions, or 3D domain swaps may have substantially altered the folds of homologous proteins [Grishin, 2001a; Andreeva et al., 2007; Alva et al., 2008], often leading to the variant form resembling an unrelated fold. In such cases, structural convergence can give the impression of continuity in an evolutionarily discontinuous landscape. We have therefore revisited the mapping of protein fold space using only homologous criteria, that is, sequence similarity.

3.2.2 Our approach

To obtain a high-level view of distant evolutionary relationships between proteins folds, we decided to cluster domains representative of all known fold types by their pairwise sequence similarity. To assemble such domains, we chose the structural classification of proteins (SCOP) database [Andreeva et al., 2008]. As described earlier, SCOP classifies proteins hierarchically by grouping related domains into families, related families into superfamilies, structurally similar superfamilies into folds, and folds into secondary structure classes. Thus, the first two levels of the classification capture homologous relationships, whereas the last two capture analogous ones. For the purpose of this study, we filtered SCOP to a maximum of 20% sequence identity and obtained 7002 domains. At this level, all superfamilies and nearly all families are still represented, but most relationships considered homologous by SCOP have been removed. Since the events we are trying to track date back to the time of the last universal common ancestor, the signal for common ancestry is expected to be weak. In such cases, methods that compare profile HMMs have been shown to be the most sensitive at detecting sequence similarity indicative of homology. We therefore employed the state-of-the-art remote homology detection method HHsearch [Soeding, 2005], which is based on profile HMMs and has been successfully used by many groups for substantiating homologies between distantly related proteins. We made pairwise comparisons of profile HMMs for the domains obtained from SCOP and clustered them in CLANS to produce a two-dimensional map of the fold space. CLANS is an implementation of the Fruchterman-Reingold graph layout algorithm to visualize pairwise sequence similarities in either two-dimensional or three-dimensional space [Frickey and Lupas, 2004]. Each sequence is represented by a point in this higher-dimensional space, in which it moves based on the forces exerted on it by other points. The force between two points is based on the sequence similarity between the corresponding sequences; in this study, we use HHsearch P-value as an indicator of sequence similarity. While significant P-values result in large attractive forces, insignificant P-values give repulsive forces. In this way, after sufficient relaxation times, similar sequences come to lie closely together. In the resulting cluster map, domains are represented by colored dots and the brightness of the lines connecting the dots indicates the degree of sequence similarity between them.

The dots in the obtained map were colored using three different schemes to produce three different views of the fold space. To analyze the high-level nature of the different architectural classes, we produced a class map by coloring the dots according to their SCOP class (Fig. 3.1). For visualizing the extent to which homologous relationships transcend current fold and superfamily boundaries, the fold (Fig. 3.2) and superfamily (Fig. 3.3) maps were obtained by coloring the dots based on their SCOP fold and superfamily, respectively.

3.2.3 Class map

Although the clustering was done only based on sequence information, we observe that proteins of the same structural class generally converge to the same regions in the map. The structural classes recognized by SCOP are: folds consisting primarily of α-helices (all-α), folds formed mainly of β-strands (all-β), folds in which helices and strands alternate regularly (α/β), and folds with irregular mixtures of helices and strands (α+β). SCOP also recognizes a small proteins class, which comprises proteins rich in cysteine, a multidomain class, and a membrane protein class, but these do not constitute classes in an architectural sense.

The class map (Fig. 3.1) shows five large regions corresponding to the four primary classes - all-α (blue), all-β (cyan), α/β (red), and α+β (yellow) - and to the small proteins class (green). We attribute their convergence to

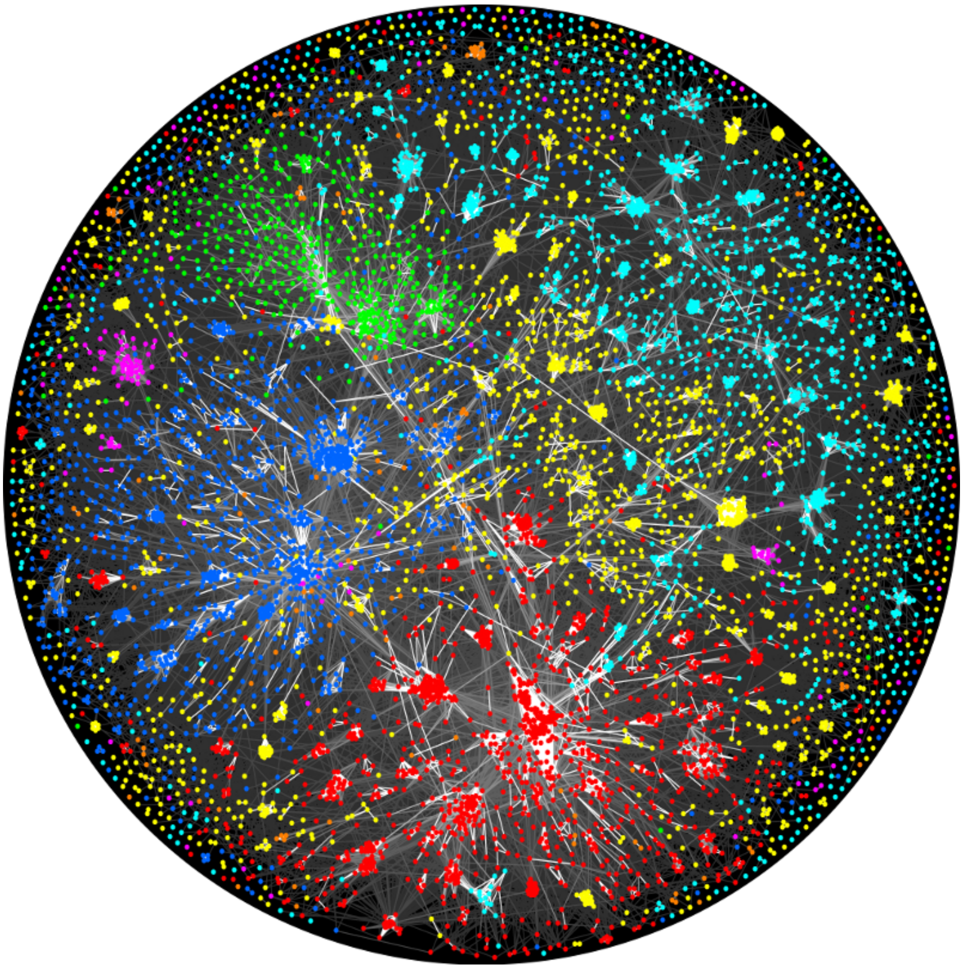


Figure 3.1: Galaxy of folds colored by classes (reproduced from [Alva et al., 2010]). Domains from the same class come to lie in similar regions of the galaxy. Domains in SCOP20 were clustered in CLANS based on their all-against-all pairwise similarities as measured by HHsearch P-values. Dots represent domains. Line coloring reflects HHsearch P-values; the brighter a line, the lower the P-value. Domains are colored according to their SCOP class: all-α (blue), all-β (cyan), α/β (red), $\alpha\&\beta$ (yellow), small proteins (green), multi-domain proteins (orange), and membrane proteins (magenta).

general similarities in amino acid composition, that is, to an analogous property. We find support for this notion in the fact that a map generated after correction for amino acid bias showed a considerably decreased grouping of the structural classes. This is consistent with previous observations that the amino acid composition reflects the structural class of a protein [Chou and Zhang, 1995]. Because of the force-directed clustering procedure, folds find their equilibrium position in the map not only by attraction to similar folds but also by repulsion of different ones. Clusters of similar folds can thus develop considerable repulsive forces, frequently clearing the areas around them and repelling dissimilar folds to distant parts of the map. For this reason, while the all-α and α/β classes are next to each other, the all-α and all-β classes occupy diagonally opposite locations. Of the primary classes, the α+β class shows the least convergence and overlaps most with the other classes, suggesting that it could be considered a catch-all class. This has already been pointed out by *Orengo et al.* [1993], who do not consider α+β a true structural class. Of the last two classes, membrane proteins cluster with the soluble proteins of the same secondary structure (helical membrane proteins with the all-α class and outer membrane proteins with the all-β class), and multidomain proteins are scattered all over the map, as their constituent domains belong to different classes.

3.2.4 Superfamily and fold maps

The current version (1.75) of the SCOP database contains 110800 domains classified into 3902 families, 1962 superfamilies, and 1195 folds. Of the 1195 folds, 136 comprise two or more superfamilies; a list of the 25 folds with the most number of superfamilies is shown in Table 3.1. We reasoned that if these superfamilies of the same fold really had multiply independent origins as highlighted by SCOP, we would obtain a map without many interconnections between them. Our map, however, shows that this is not the case. Although unrelated domains from the same class are very loosely connected in general, the many tighter clusters are formed from groups of domains with significant pairwise similarities that are indicative of homology. We chose 60 visually prominent clusters for further analysis. As expected, most of these contain domains of the same superfamily, but 18 contain domains from different superfamilies. Out of these, seven comprise superfamilies of the same fold and 11 superfamilies of different folds.

One large cluster contains the various superfamilies of the aforementioned TIM (β-α)$_8$-barrel (yellow cluster at the bottom in Fig. 3.2). In our map, all 33 superfamilies of this fold (SCOP c.1.1-c.1.33), except for monomethylamine methyltransferase (c.1.25) and NAD(P)-linked oxidoreductase (c.1.7), cluster into three groups, which are tightly linked to each other, in agreement with their proposed homology [Copley and Bork, 2000; Nagano et al., 2002; Soeding, 2005]. Other examples of such folds with tightly connected superfamilies include the α/α toroid fold (a.102, salmon cluster near the left edge in Fig. 3.2) and the β-trefoil fold (b.42). In both of these cases, a homologous origin for the superfamilies within the fold is likely [Ponting and Russell, 2000; Liang et al., 2002].

Although our results indicate that folds may not be as polyphyletic as assumed by SCOP, we do see instances of analogous folds. The most striking example is the ferredoxin-like fold (d.58), which has a two-layered α + β

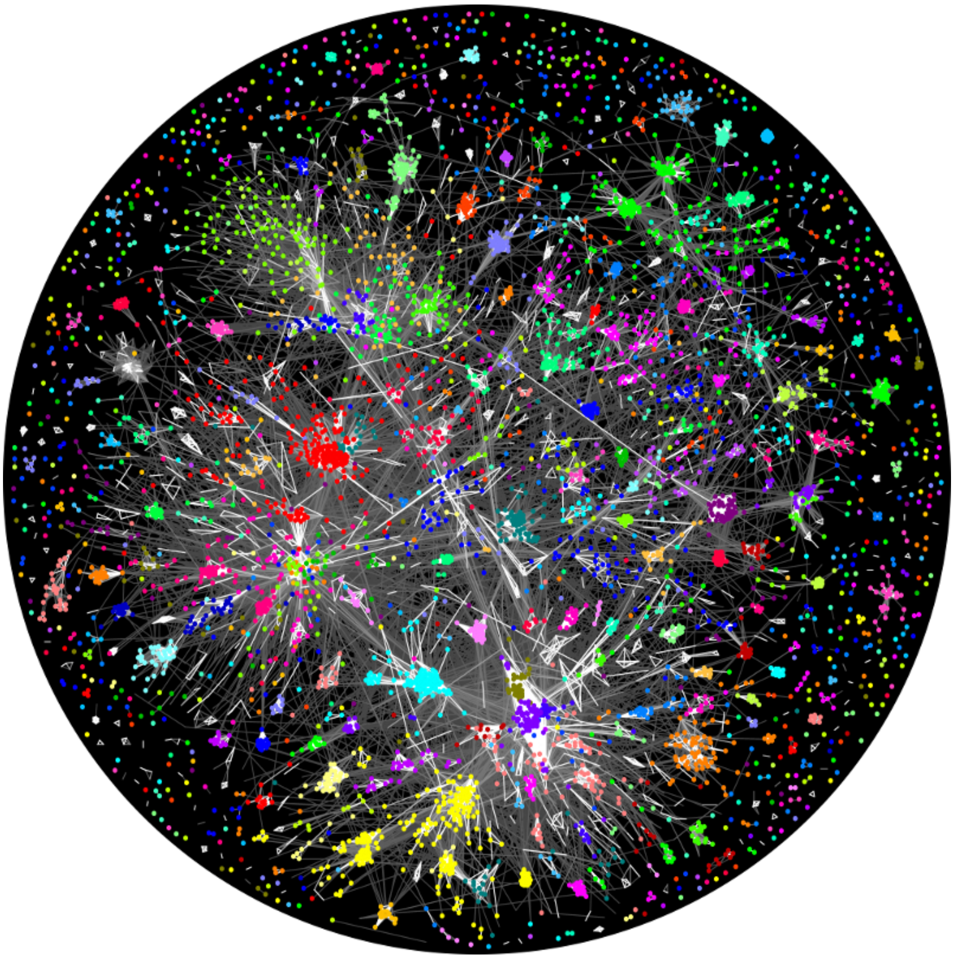


Figure 3.2: Galaxy of folds colored by folds (reproduced from [Alva et al., 2010]). Some clusters connect domains of different fold, pointing to common, homologous fragments of similar sequence and structure. These might represent descendants of a set of ancient peptide modules, from which the first protein domains have been assembled

architecture, comprising two α-helices and four β-strands, and has by far the largest number of superfamilies in SCOP, 59 in total (Table 3.1). These superfamilies are distributed all over the map, indicating that they converged upon the same fold independently. Other examples are the ferritin-like folds (a.25), the spectrin repeat-like folds (a.7), four-helical up-and-down bundle folds (a.24), and the bromodomain-like folds (a.29). We also see instances of superfamilies of the same fold that show a mixture of homologous and analogous connections. Examples include the SH3-like barrel fold (b.34), the OB-fold (b.40), the β-Grasp fold (d.15), and the DNA/RNA-binding three-helical bundle fold (a.4). Of the 25 folds with the most number of

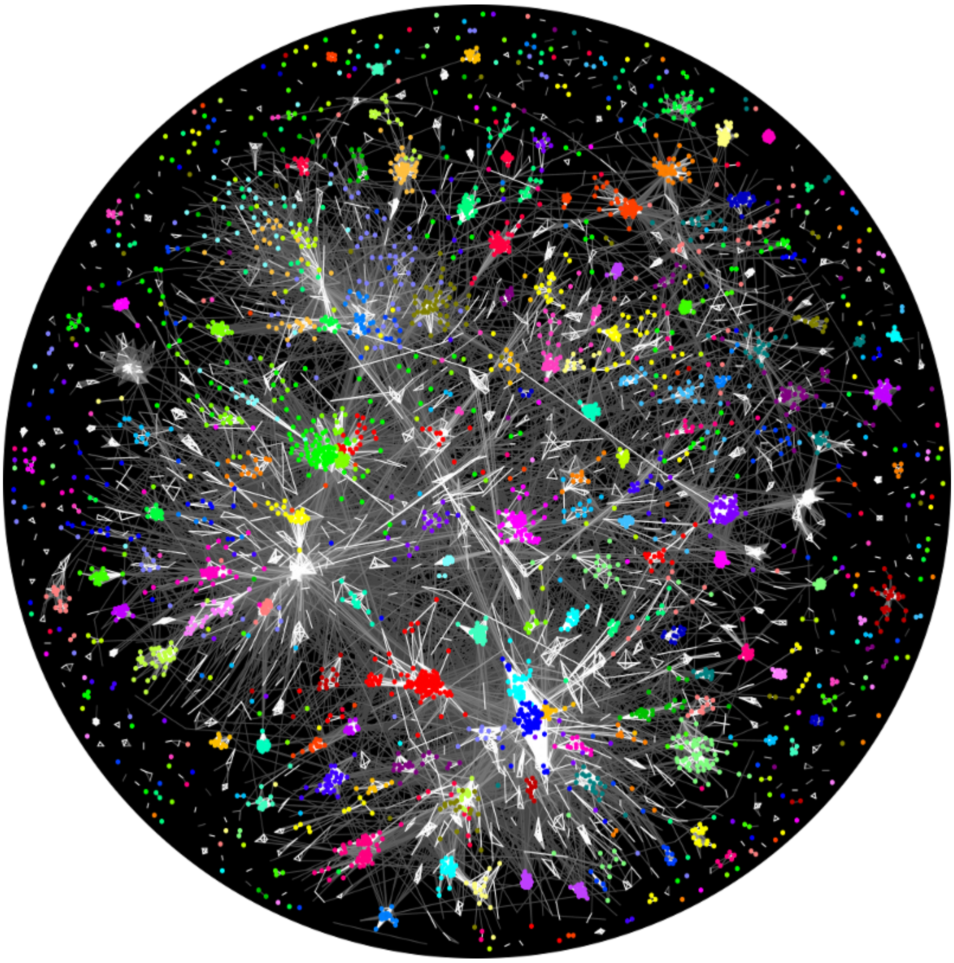


Figure 3.3: Galaxy of folds colored by superfamilies (reproduced from [Alva et al., 2010]). Many tight clusters contain various superfamilies of the same fold, indicating that folds with multiple independent origins are rather the exception than the rule.

superfamilies in SCOP, nine folds comprise of analogous superfamilies, seven contain homologous superfamilies, and nine folds have superfamilies that show a mixture of analogous and homologous connections (Table 3.1). SCOP does not consider one of these 25 folds, the single transmembrane helix fold (f.23), a true fold, as it is a catch-all fold for unrelated single transmemebrane helices. In line with this, the various superfamiles of this fold are scattered all over the map.

Of the 11 clusters comprising domains belonging to different folds, connections in two clusters rely on global similarities between domains. One cluster contains β-propellers, which are toroidal folds with between four and ten repeats of a four-stranded β-meander. In SCOP, they are classified into

Table 3.1: 25 SCOP folds with the most number of superfamilies

SCOP Fold	Fold name	Number of superfamilies	Remarks
d.58	ferredoxin-like	59	mostly analogous
f.23	single transmembrane helix	38	not a true fold analogous
c.1	TIM $(\beta/\alpha)_8$-barrel	33	mostly homologous
b.1	immunoglobulin-like	28	mixed
a.24	four-helical up-and-down bundle	27	mostly analogous
a.118	α-α superhelix	24	mixed
b.34	SH3-like barrel	21	mixed
a.2	long α-hairpin	20	mixed
g.3	knottins (small inhibitors, toxins, lectins)	19	mixed
g.41	rubredoxin-like	17	mostly homologous
a.7	spectrin repeat-like	16	mostly analogous
a.60	SAM domain-like	16	mostly homologous
b.40	OB-fold	16	mixed
a.29	bromodomain-like	15	mostly analogous
c.23	flavodoxin-like	15	mostly homologous
d.15	β-Grasp	14	mixed
b.69	seven-bladed β-propeller	14	mostly homologous
a.4	DNA/RNA-binding three-helical bundle	14	mixed
a.137	non-globular all-α subunits of globular proteins	14	mostly analogous
b.68	six-bladed β-propeller	11	mostly homologous
d.129	TBP-like	11	mixed
a.8	immunoglobulin/albumin-binding domain-like	11	mostly analogous
d.52	α-lytic protease prodomain-like	10	mostly analogous
d.110	profilin-like	10	mostly homologous
b.2	common fold of diphtheria toxin/transcription factors/ cytochrome f	10	mostly analogous

five different folds (b.66-b.70), each with multiple superfamilies. Recently, it was proposed that all β-propellers have a common origin [Chaudhuri et al., 2008], and we find that they indeed cluster together, except for apyrase (b.67.3) and sema domain (b.69.12), which contain large insertions. The second cluster comprises transmembrane β-barrels, which are classified into seven superfamilies within two folds (f.4 and d.24.1.4) in SCOP. Their homologous origin has been discussed recently [Arnold et al., 2007; Remmert et al., 2009].

In the remaining nine clusters, the connections between domains clearly result from the presence of sequence- and structure-similar subdomain-sized fragments. For example, one large cluster contains a variety of topologically distinct DNA-binding domains with a common helix-turn-helix motif (large, mainly red cluster at the middle in Fig. 3.2), whose homologous origin has been discussed previously [Brennan, 1993]. In SCOP, these domains are classified into 16 superfamilies contained within 10 folds. Another large cluster contains the Rossmann folds (large cluster comprising dots with various colors at the bottom in Fig. 3.2), which possess a common dinucleotide-binding β-α-β-element. Their evolutionary relationship has also been proposed previously [Lupas et al., 2001; Xie and Bourne, 2008]. A further cluster comprises the eukaryotic (type-I) and the prokaryotic (type-II) KH-domains, which are topologically distinct but homologous [Grishin, 2001b]. The similarity between these folds is limited to a α-α-β motif. These clusters lend support to a theory on the origin of folded proteins, which proposes that these structure- and sequence-similar fragments seen in disparate molecular contexts represent remnants of an ancient peptide-RNA world, thus suggesting that today's domains have arisen by fusion, amplification, and divergence from a simpler set of peptide modules [Fetrow and Godzik, 1998; Lupas et al., 2001; Soeding and Lupas, 2003]. In the next section, we present 50 such fragments that we found following a systematic search.

3.2.5 Discussion

Our two-dimensional map of domains of known structure offers a global view of evolutionary relationships in the fold space and shows the preponderance of homologous connections that transcend the superfamily level, revealing that many folds are monophyletic rather than have independent origin. While many of the relationships observed in the map have been discussed individually before, confirming the validity of these findings, we do not observe some clusters that we expected from reported instances of remote homology. One reason is that many of these involve domains with few homologs of known structure. For clustering, these would have to rely entirely on the strength of their pairwise connection rather than benefiting from the stronger attractive field generated by a compact group of homologous domain families. Another reason is that some domains that are clearly recognizable at the structural level do not appear as independent entries in SCOP. The homology between the histone fold (a.22) and the C-domain of AAA+ ATPases (c.37.1.20) that we detected in another study is a case in point (see Chapter 4). In the present map, although domains belonging to these two folds show clear pairwise connections, they do not form a tight cluster. This is because C-domains are not characterized as a separate fold in SCOP but are classified with other P-loop NTPases based on the preceding ATPase domain; they therefore cluster tightly with these. We also anticipate that some instances of distant homology remain unrecognized in our map if they involve domains with few homologs in current databases, as it is not possible to build a reasonable profile HMM in these cases. With the progress of sequencing projects, this problem should wane. A few links with significant HHsearch P-values are false positives, which connect clearly unrelated domains, such as the link between many TIM barrel proteins and the guanine deaminase (d2ooda1, turquoise cluster at the bottom in Fig. 3.2). According to a systematic analysis of the highest-scoring false positives, the chief cause for these false links are corrupted alignments that are used to build the profile HMMs. In this case, sequences from TIM barrels have crept into the alignment of d2ooda1 during the iterative search.

3.2.6 Materials and methods

We used the SCOP database, version 1.75, filtered to a maximum of 20% sequence identity (SCOP20) as provided by ASTRAL [Brenner et al., 2000]. HHsearch was used for all-against-all comparison of the 7002 domains in SCOP20. Multiple alignments were built for each of the SCOP20 domains using the buildali.pl script (with default parameters) from the HHsearch 1.6.0 package. This script uses PSI-BLAST [Altschul et al., 1997] and contains heuristics to reduce the inclusion of nonhomologous sequence segments at the ends of PSI-BLAST sequence matches, the leading cause of high-scoring false positive matches in PSI-BLAST. Profile HMMs were calculated from the alignments using hhmake and compared with HHsearch, both from the HHsearch 1.6.0 package. We switched off secondary structure scoring and the compositional bias correction (options -ssm 0 -sc 0) and used default settings otherwise. We clustered the SCOP20 domains by their pairwise HHsearch P-values in CLANS [Frickey and Lupas, 2004], an implementation of the Fruchterman- Reingold clustering algorithm that scales negative log-P-values into attractive forces in a force field. Clustering was done to equilibrium in 2D at a P-value cutoff of 1.0e-01 using default settings.

3.3 Reconstructing the observable remnants of the RNA-peptide world

3.3.1 Proteins from peptides

Although it is largely recognized that the diversity of modern proteins originated by mutation, duplication, and shuffling from a limited set of ancestral domains, most of which were already established at the time of the LUCA, the origin of this limited set itself is poorly understood. As explained earlier (Chapter 2), it is highly improbable that nature built the first domains from scratch; but if domains were not built *de novo*, how did they emerge? The detection of numerous examples of protein domains composed of multiple copies of a homologous repeat, together with the growing number of examples of sequence- and structure-similar fragments within nonhomologous protein folds suggests that domains themselves might be divisible and have arisen from smaller fragments [Fetrow and Godzik, 1998; Lupas et al., 2001; Soeding and Lupas, 2003]. It is therefore attractive to think that the first domains themselves arose from a limited set of proto-domains, much in the same way as modern multidomain proteins evolved from these domains. We are following the hypothesis that folded domains arose by fusion and recombination from a simpler, ancestral set of peptides, which originally served as cofactors of an RNA world (see Chapter 2). If this hypothesis is true, we would expect to see remnants of these peptides in modern proteins and systematic comparisons should allow us to reconstruct this ancient set in a manner similar to how linguists have reconstructed ancient vocabularies by comparing modern languages.

Today, for instance, it is recognized that the archaic Greek word *polis*[1] (plural *poleis*), which meant city-states in ancient Greece, is the ancestor of hundreds of words in many modern European languages (Fig. 3.4). Unlike other ancient city-states of the time that were ruled by kings, *poleis* were governed by a council comprising of its own inhabitants. Over time, these city-states grew bigger by incorporating neighboring regions and the notion of citizenship came into being. This development also changed the meaning of the word *polis* from describing city-states to describing the entire body of citizens of a city-state. The idea of *polis* and the derivatives of the word *polis* itself made their way into other civilizations and languages of the time. One such derivative in Latin, *politia*, meaning state, in turn gave rise to many words in many languages; for instance, descendants in English include police, policy, polity, and politics. The root *polis* is also seen in the names of cities around the world, e.g. Persepolis in Iran, Nicopolis in Palestine, Sozopol in Bulgaria, Stavropol in Russia, and Napoli in Italy, and in endings of words that refer to a special kind of place, e.g. Necropolis (city of the dead), Acropolis (high city), and Pentapolis (a group of five cities). Furthermore, it is believed that the Greek root *polis*, the Lithuanian *pilis*, meaning castle, the Latvian *pils*, meaning castle, and the Sanskrit *pura*, meaning city, are homologous, and arose from a Proto-Indo-European root *pel-*, meaning full. The Sanskrit root *pura*, for instance, is seen in the names of cities in Asia, e.g. Jaipur (India) and Singapore. This example illustrates how an useful ancient root has extensively spread to several languages across the globe. While many of the ancestral derivatives are not in use anymore, we can still make reliable reconstructions of the evolutionary path through which the archaic word *polis* gave rise to the words we know today; this is done by comparing modern languages. In an analogous manner, we wish to identify ancient peptides, which were used in different contexts in the evolution of

[1] Information provided in this paragraph was compiled from the following sources: `http://en.wikipedia.org/wiki/Polis`, `http://en.wiktionary.org/wiki/polis`, and `http://www.etymonline.com/`.

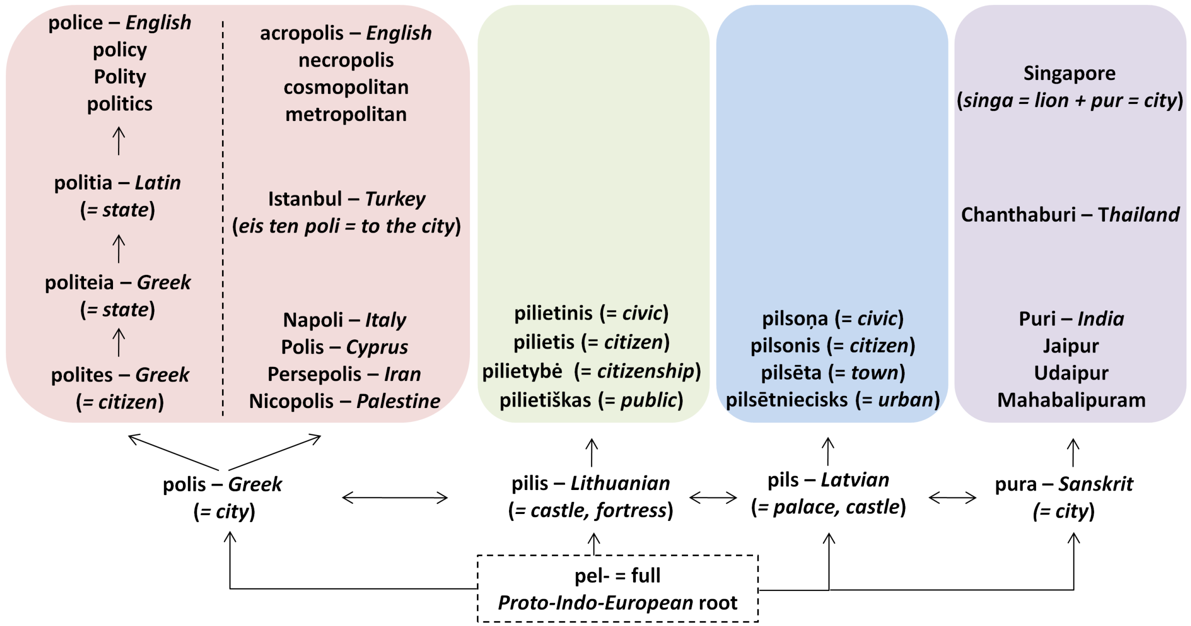


Figure 3.4: The archaic Greek word *polis* and its cognates. Today, the ancient Greek word *polis* (meaning city) is recognized as the ancestor of many words in many modern European languages, e.g. derivatives of *polis* in English include police, policy, and politics. It also appears in the endings of many words that describe a special kind of place (e.g. necropolis, meaning city of the dead) and in the names of cities across the world (e.g. Napoli). In some modern words, the signal for origin from *polis* is very weak, e.g. Istanbul, the largest city in Turkey, is a corruption of the Greek phrase *ein ten poli* meaning *into the city.* Cognates of *polis* include Lithuanian *pilis*, Latvian *pilis*, and Sanskrit *pura*, all three meaning town or fort. The Sankrit root *pura* is seen in the names of several cities in India (e.g. Jaipur) and south Asia (e.g. Singapore). It is thought that these three roots originate from a proto-Indo-European root pel-, meaning full.

folded proteins, by comparing modern proteins.

3.3.2 Previous studies

Most hitherto studies on the origin of domains focused either on individual cases of repeats found within a domain or on individual instances of structure- and sequence-similar fragments that co-occur in seemingly unrelated domains. The earliest of these studies were conducted in the 1970s by Andrew McLachlan, who described many examples of evolution of domains by repetition, and put forward the idea that repetition is not only an important mechanism in the evolution of multidomain proteins, but also in the evolution of individual domains themselves [McLachlan, 1979; McLachlan et al., 1980; McLachlan, 1980]. He based this hypothesis primarily on internal symmetries observed in folds at the structural level, a property that typically reflects analogy; however, a number of studies have since presented many examples of folds that show internal symmetry on the structural as well as sequence level, indicating that repetition is indeed a mechanism for the origin of domains. Many highly populated folds, such as TIM barrels [Copley and Bork, 2000; Nagano et al., 2002; Soeding, 2005], β-propellers [Chaudhuri et al., 2008], and ferredoxins [Otaka and Ooi, 1987], show internal symmetries both at the sequence- and structure-level, and are therefore believed to have arisen through amplification from ancestral units. Such internal symmetry is also observed in a large class of membrane-embedded domains, the outer membrane β-barrels [Remmert et al., 2009], which comprise between four and twelve β-β hairpins in a circular arrangement.

In addition to illuminating on folds composed of homologous repeats, some studies also described examples of local sequence and structure similarities in domains with different folds. The most noteworthy cases include the KH (α-α-β) motif, which occurs in type-I and type-II KH domains that otherwise assume different folds [Grishin, 2001b] and the helix-hairpin-helix motif, which is common to a number of topologically distinct nucleic acid-binding domains [Doherty et al., 1996]. These local similarities were thought to have originated either due to domains with different evolutionary origins having converged on similar local motifs or due to a proto-domain being decorated in different ways, leading to folds that look different otherwise. Spurred by these individual cases of local similarities, *Lupas et al.* [2001] performed a systematic comparison of domains of known fold types using a structure- and sequence-based approach and detected seven examples of sequence- and structure-similar motifs that occur in different folds. Based on the presence of these locally similar motifs and on the numerous examples of domains containing multiple copies of a homologous repeat, they hypothesized that the diversity of modern proteins may have arisen from peptide ancestors, referred to as antecedent domain segments (ADSs), which first emerged in the RNA world. From this perspective, the local similarities observed in modern domains might represent the visible remnants of such ancient peptides. To the best of our knowledge, no study, since the study of *Lupas et al.* [2001], has attempted to systematically detect ADSs.

3.3.3 Our approach

We decided to revisit this topic and systematically search for ADSs for many reasons. First, the recent growth of protein sequence and structure databases have led to a substantial expansion in the evolutionary depth of domain families and have provided structural representatives for a large number of domain families. Furthermore, this growth has also brought about a considerable increase in the number of domain superfamilies and fold types. We reasoned that this wide availability of sequence and structure data would allow us to detect more ADSs.

Second, the approach employed by *Lupas et al.* [2001] for detecting ADSs had some limitations. For a given a pair of protein structures, their method first detected contiguous motifs that were common to both structures and contained similar side-chain patterns. These motifs were then subjected to statistical significance tests to filter out motifs that were similar due to structural constraints, rather than because of divergence from a common ancestor. They, however, ignored motifs with conserved side-chain patterns arising from amino acids unlikely to be directly involved in molecular function in their searches. By excluding these amino acids (Ale, Phe, Ile, Leu, Pro, and Pro), the search space was reduced significantly, possibly resulting in many potential ADSs not being detected. For instance, this approach, as also admitted by the authors, fails to detect motifs with hydrophobic residue conservation, examples of which include the helix-hairpin-helix motif [Doherty et al., 1996] and the KH motif [Grishin, 2001b]. They also restricted their search to motifs containing up to 20 amino acid residues, leading to potential ADSs of lengths more than 20 residues being ignored.

Finally, we were also propelled by the availability of a sensitive remote homology detection method, HHsearch that compares profile HMMs with each other [Soeding, 2005]. Profile HMMs encode position-specific amino acid and gap preferences that are conserved for much larger evolutionary distances than amino acid identities. HHsearch is currently one of the most sensitive and widely accepted remote homology detection methods, and has been successfully used, by us and others, for inferring distant homologies, especially for ones that date back to the time of the LUCA. It has also been employed to detect homologous fragments that occur in domains with different folds. For example, in a recent study HHsearch was used to infer homology between the different superfamilies of outer membrane β-barrels (OMBBs), which are a major class of outer membrane proteins (OMPs) in gram-negative bacteria, mitochondria, and chloroplasts. They consist of 4-12 β-β hairpin repeats that form a closed β- barrel around a central pore. Although the β-β hairpins from within a OMBB or from different OMBBs are structurally well superposable, they do not show clear sequence similarity to each other, and thus a scenario for an origin by amplification or for a common ancestry had not been previously proposed. *Remmert et al.* [2009] employed HHsearch to show that all OMBB superfamilies exhibit sequence similarity indicative of homology. They also detected sequence repeats that coincide with the β-β hairpins in most OMBBs and based on this they proposed that OMBBs arose by the amplification of an ancestral β-β module.

To reconstruct the set of putative ancient peptides, we aimed at finding sequence- and structure-similar fragments that reoccur in domains of different folds types, i.e. in domain currently considered to be analogous. To assemble domains representative of all known fold types, we chose the SCOP database [Andreeva et al., 2008] and filtered it to a maximum of 20% sequence identity and obtained 7002 domains. At this level, all folds, superfamilies, and nearly all families are still represented, but most relationships considered homologous by SCOP have been removed.

Since the events that led to the emergence of domains took place much before the time of the LUCA, modern domains may only retain weak signals of these events in their sequences. Since structures diverge much more slowly, their similarity is often used to identify such distant events. However, similar structures may have arisen convergently from different origins, owing to the limited number of structural solutions available to a folded polypeptide chain, and structure similarity thus frequently does not provide conclusive evidence of common ancestry. In contrast to structure space, the combinatorial sequence space is vast and many sequences are compatible with a particular local structure, so that sequence convergence is rare. Thus sequence similarity is accepted as the most important marker for common ancestry. Therefore, we have used sequence similarity, as evaluated by the aforementioned remote homology detection method HHsearch, for inferring common ancestry of domains in this study.

We employed a sequence- and structure-based approach to detect homologous fragments in the SCOP20 set. Since we were primarily interested in inter-fold relationships, we only compared domains of different folds with each other using HHsearch. We filtered out all pairwise matches with a HHsearch probability of below 50%, to obtain domains of different folds that show significant similarity in sequence. HHsearch probability is an estimate for the likeliness of a given match to be a true positive. In many previous studies, we have observed that at probability values of above 50%, the error rate of HHpred is extremely low (see for example [Kopec et al., 2010] and [Alva et al., 2007]), and we therefore chose this cut-off for the current study. Subsequently, we calculated structural superpositions for the pairwise sequence segments aligned by HHsearch and removed all matches with a RMSD of worse than 2Å, to obtain pairwise matches that are similar both in sequence and structure. We then eliminated all matches in which connections relied on fragments comprising less than eight amino acids. This was done to filter out short and single secondary structural elements because such simple elements were invented independently multiple times over the course of protein evolution and as a result inferring homology between them is difficult. This filtered set was then used to extract groups of domains of different fold, whose similarity were hinged on the presence of shared sequence- and structure-similar fragments.

3.3.4 Results and discussion

After applying our approach, we obtained 655 clusters, each comprising pairwise matches between domains of different folds, which were based on the presence of a shared sequence- and structure-similar segment (see Table 3.2 and Table 3.3 for examples). 503 of these clusters comprised only a single pairwise match. Many of these hits were between isolated members of superfamilies, whose other members did not make connections to each other. This suggested that the detected relationships were possibly false positives. In all such cases, we evaluated the validity of the matches by comparing additional domains from a bigger set, which contained SCOP domains clustered down to a pairwise sequence identity of 40% (SCOP40). Clusters in which no further support for the detected relationship was found were filtered out. The remaining clusters with a single pairwise match included either connections between superfamilies that contain just one domain in SCOP20 or comprised matches that were also part of other larger clusters and were not merged together because of our stringent clustering criteria. These clusters were analyzed further on a case-by-case basis.

Of the clusters that contained two or more pairwise matches, the connections in some relied on the global similarity of the folds, rather than on the presence of sequence- and structure-similar fragments. For instance, one of the clusters contained domains of the ‘ribosomal proteins S24e, L23, and L15’ fold (SCOP d.12; e.g. d1vqos1) and of the RNA-binding domain superfamily of the ferredoxin-like fold (d.58.7; d2ghpa2), which, despite of being classified into different folds, are clearly related by a circular permutation event. We filtered out all clusters in which connections between domains were due to the global similarity of their folds. In many other clusters, connections between domains were based on fragments that were entirely contained within aligned segments of other clusters. This, for example, was the case with β-propellers (folds b.66 to b.70), which, as described earlier, are toroidal folds with between four and ten homologous repeats of a four-stranded β-meander [Chaudhuri et al., 2008]. The matches between domains of these folds were contained in 15 different clusters. We attribute this to the fact that the aligned segments of β-propellers of different folds were frequently much shorter or much longer than the four-stranded β-meander

domain-1 accession	SCOP ID	domain-2 accession	SCOP ID	HHsearch probability	RMSD (Å)
d1tuaa1 (4−34)	d.51.1.1	d2uubc1 (62−92)	d.52.3.1	89.2	1.47
d1tuaa1 (7−37)	d.51.1.1	d1wf3a2 (54−85)	d.52.3.1	88.0	0.86
d2ctfa1 (23−51)	d.51.1.1	d1wh9a_ (37−65)	d.52.3.1	80.4	1.94
d2je6i3 (2−37)	d.51.1.1	d1wh9a_ (36−71)	d.52.3.1	91.2	1.82
d1zzka1 (5−34)	d.51.1.1	d2uubc1 (62−91)	d.52.3.1	86.3	1.64
d2cpqa1 (10−37)	d.51.1.1	d2uubc1 (63−90)	d.52.3.1	88.1	1.99
d2ctma1 (12−41)	d.51.1.1	d2uubc1 (63−92)	d.52.3.1	86.8	1.90
d2je6i3 (2−32)	d.51.1.1	d2uubc1 (62−92)	d.52.3.1	89.2	1.65
d1zzka1 (5−37)	d.51.1.1	d1wf3a2 (51−84)	d.52.3.1	85.8	1.83
d2je6i3 (3−38)	d.51.1.1	d2asba3 (28−63)	d.52.3.1	96.1	1.73
d2ctma1 (9−48)	d.51.1.1	d2asba3 (25−64)	d.52.3.1	95.7	1.90
d2ctma1 (10−41)	d.51.1.1	d1wf3a2 (50−82)	d.52.3.1	90.8	1.84

Table 3.2: Cluster comprising pairwise matches between type-I (d.51.1) and type-II (d.52.3) KH-domains. These two domains are topologically distinct, but share a homologous α-α-β-motif (Fig. 3.5, fragment 11), which is involved in binding RNA or single-stranded DNA.

domain-1 accession	SCOP ID	domain-2 accession	SCOP ID	HHsearch probability	RMSD (Å)
d1tzya_ (47-87)	a.22.1.1	d1qvra1 (7-49)	a.174.1.1	85.2	1.76
d1tzya_ (48-77)	a.22.1.1	d1w5sa2 (257-286)	c.37.1.20	55.9	1.33
d1jfia_ (37-65)	a.22.1.3	d1qvra1 (7-35)	a.174.1.1	52.4	1.00
d1fnna2 (244-272)	c.37.1.20	d1tzya_ (49-77)	a.22.1.1	54.9	0.97
d1tafa_ (35-64)	a.22.1.3	d1w5sa2 (256-285)	c.37.1.20	67.0	1.09
d1w5sa2 (257-286)	c.37.1.20	d1q9ca_ (134-163)	a.22.1.3	56.3	1.38
d1jfia_ (39-66)	a.22.1.3	d1fnna2 (244-271)	c.37.1.20	71.6	0.81
d1jfia_ (37-65)	a.22.1.3	d1k6ka_ (6-34)	a.174.1.1	61.9	1.55
d2huec1 (45-74)	a.22.1.1	d1fnna2 (244-273)	c.37.1.20	80.2	0.83
d1f1ea_ (114-145)	a.22.1.2	d1fnna2 (243-274)	c.37.1.20	62.8	0.81
d1n1ja_ (41-71)	a.22.1.3	d1fnna2 (245-275)	c.37.1.20	74.5	0.67
d1fnna2 (243-272)	c.37.1.20	d1tafa_ (36-65)	a.22.1.3	61.5	0.72
d1q9ca_ (133-166)	a.22.1.3	d1k6ka_ (84-117)	a.174.1.1	72.7	1.76
d1k6ka_ (7-36)	a.174.1.1	d1n1ja_ (39-68)	a.22.1.3	52.4	1.72

Table 3.3: Cluster comprising pairwise matches between histones (a.22.1), the helical part of the extended ATPase domains found in AAA+ proteins (the C-domain; c.37.1.20), and the N-terminal domain of Clp/Hsp100 proteins (Clp N-domain; a.174.1). The connections between these domains is based on the presence of a shared helix-strand-helix motif (Fig. 3.5, fragment 9).

repeat, and because of this our clustering procedure failed to group them together. In all such cases, we investigated the relationships further by looking at the relative position of the domains in consideration in the galaxy of folds (Fig. 3.2 and Fig. 3.3), and also by performing additional HHsearch searches with domains from the SCOP40 set. Most superfamilies of the five β-propeller folds, for instance, cluster together in the galaxy of folds, suggesting that they are evolutionarily related. We therefore merged the 15 different clusters comprising fragments from β-propellers together.

All clusters in which the connections relied on the presence of a shared sequence- and structure-similar fragment were examined on a case-by-case basis. The detected relationships were evaluated further by performing additional comparisons with domains from the SCOP40 set and by generating sequence alignments and structural superpositions, and visualizing them them interactively. After looking through all clusters, we found 50 sequence- and structure-similar fragments (potential antecedent domain segments), each occurring in domains that belong to two or more different SCOP folds (Fig. 3.5; detailed information on each fragment is provided in Appendix A). The median length of these fragments is 24 residues, with the shortest fragment comprising 9 residues and the longest one 60 residues and, on average, they contain two or three secondary structure elements (i.e. α-helices and β-strands). This is in accord with our expectation that ancient peptides were simple and subdomain-sized. About a third (17) of these 50 fragments have been described individually before. This, to our knowledge, includes all previously reported cases of sequence- and structure-fragments, and thus confirms the validity of our approach. Examples of such fragments include the KH motif (Fig. 3.5, fragment 11), common to the type I and type II KH domains [Grishin, 2001b], the helix-hairpin-helix motif (Fig. 3.5, 2), found in a number of different nucleic acid-binding domains [Doherty et al., 1996], the ASP box (Fig. 3.5, 36), seen in carbohydrate-binding domains belonging to six different folds [Copley et al., 2001], the EF-Tu-binding α-hairpin (Fig. 3.5, 6), found in elongation factor EF-Ts and the ribosomal protein L7/12 [Wieden et al., 2001], and the P-loop (Fig. 3.5, 22), shared by PEP carboxykinase-like and P-loop containing nucleoside triphosphate hydrolase domains [Gay and Walker, 1983; Matte et al., 1996]. For many of these previously described fragments, we identified new structural contexts. For example, in addition to the two known contexts of the aforementioned P-loop (Fig. 3.5, 22), we found a third one, namely the catalytic domain of MurD-like peptide ligases (SCOP c.72.2); like in the other two domains, the P-loop is, in this case too, involved in binding the phosphate backbone of a mononucleotide.

The 50 homologous fragments we found are spread across a total of about 162 folds and 225 superfamilies in SCOP. Given that the current version of SCOP comprises 1195 folds and 1962 superfamilies, our fragments represent a considerable fraction of all folds (13%) and superfamilies (11%) in SCOP.

Furthermore, of the 25 most populated folds in SCOP, which comprise about 25% of all superfamilies (Table 3.1), 19 folds contain superfamilies that encompass at least one of these 50 fragments[2]. This makes clear that, far from being anecdotal, homologous (sequence- and structure-similar) fragments are quite widespread in modern domains and thus provides strong evidence for our proposition that domains arose from simpler subdomain-sized fragments. 18 of these 50 fragments are found in three or more folds, whereas the rest are seen in two folds each. This is analogous to the situation of domains in multidomain proteins: while some domain types are widespread in proteins, the majority of domains are only seen in a few contexts. The most widespread of these fragments, the DNA-binding helix-turn-helix motif (Fig. 3.5, fragment 1), is found in 18 folds comprising 24 superfamilies. Other widespread fragments include the dinucleotide-binding β-α-β-motif (10 folds/10 superfamilies; Fig. 3.5, fragment 21), the DNA-binding helix-hairpin-helix motif (8/15; fragment 2), the TPR element (8/16; fragment 31), and the GD-box-containing β-α-β motif (6/8; fragment 41).

Based on many lines of evidence, we propose that the 50 fragments we identified may represent the observable remnants of the RNA-peptide world from which folded proteins emerged. First, more than a third (18) of these fragments are either directly involved in interactions with nucleic acids or are found in nucleic acid-binding domains (Fig. 3.5, fragments with a yellow background), including domains of the ribosomal proteins, many of which are ubiquitous to all life forms on earth and are thought to have been established already at the time of the LUCA. Examples of DNA-binding motifs include the helix-turn-helix motif (Fig. 3.5, fragment 1), the helix-hairpin-helix motif (fragment 2), and the helix-strand-helix motif (fragment 9), and that of RNA-binding motifs include the KH-motif (fragment 11) and the α-L-motif (fragment 15). According to the ADS theory, in the pre-biotic world peptides were initially optimized to become structured using RNA as scaffolds and then the increase in complexity resulting from the fusion of the first peptides yielded folding as an emergent property [Lupas et al., 2001]. From this perspective, nucleic acid-binding domains are possibly among the most ancient domains. It thus seems plausible to assume that the high incidence of nucleic acid-binding motifs in our set of ancient peptides highlights the ancestral nature of this set. In fact, for some of these fragments there is evidence to suggest that they were present in the LUCA. For instance, the DNA/RNA-binding three-helical bundle fold (a.4), which is one of the 18 folds that contains the DNA-binding helix-turn-helix motif (fragment 1), is omnipresent in organisms from diverse branches of life and is thought to be one of the five most ancient folds [Caetano-Anolls et al., 2007].

[2] The 19 folds are: a.2, a.4, a.7, a.8, a.24, a.60, a.118, b.1, b.2, b.34, b.40, b.68, b.69, d.15, d.52, d.58, d.110, g.3, and g.41.

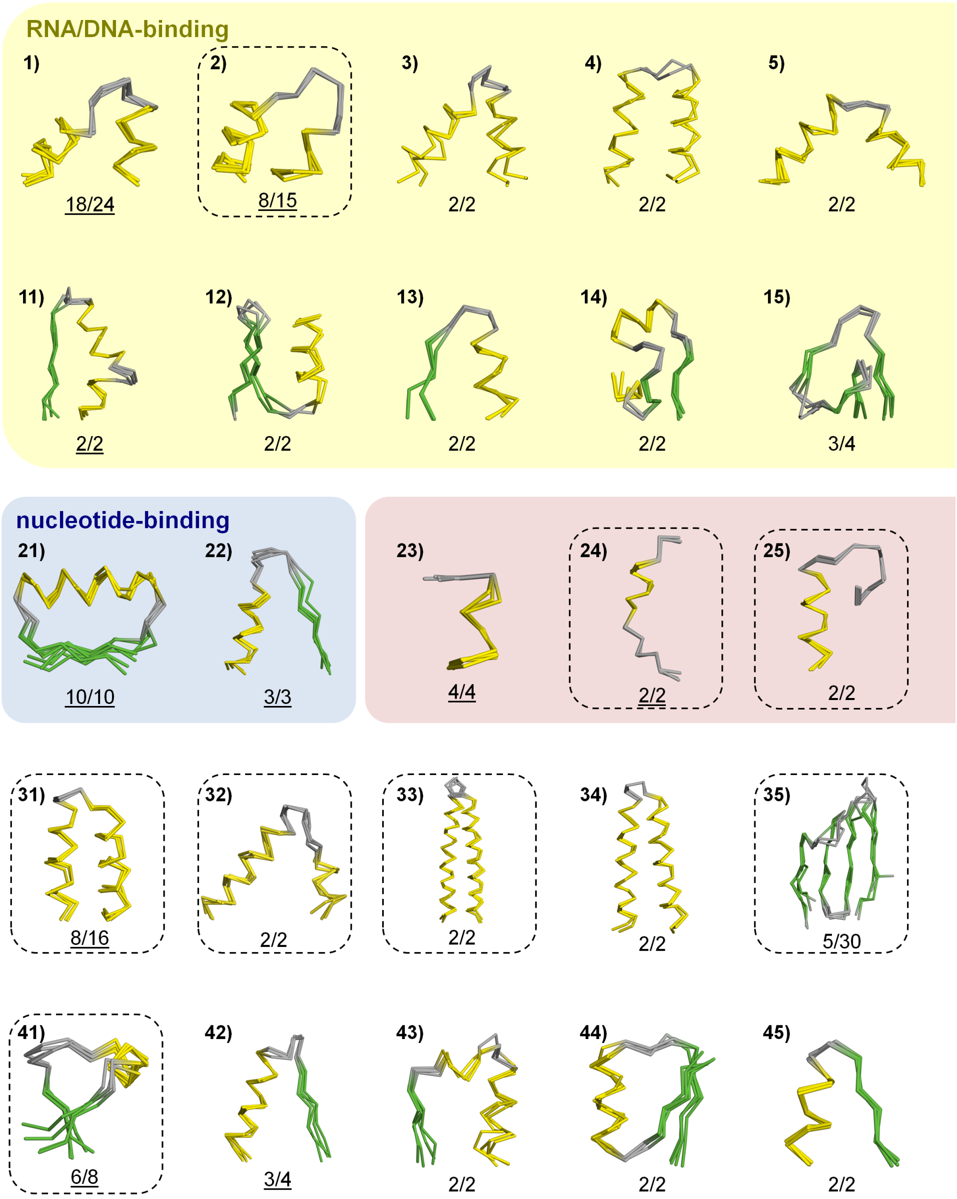


Figure 3.5: Dictionary of ancient peptides. We compared domains of known fold types using a sequence- and structure-based approach and detected 50 fragments that occur in domains of different folds, but are still similar in sequence and structure. This suggests that they might represent the observable remnants of the peptides from which the first folded domains arose. Each of the 50 fragments is shown in backbone representation; α-helices are colored in yellow, β-strands in green, and loops in gray. Detailed information on each fragment is provided in Appendix A.

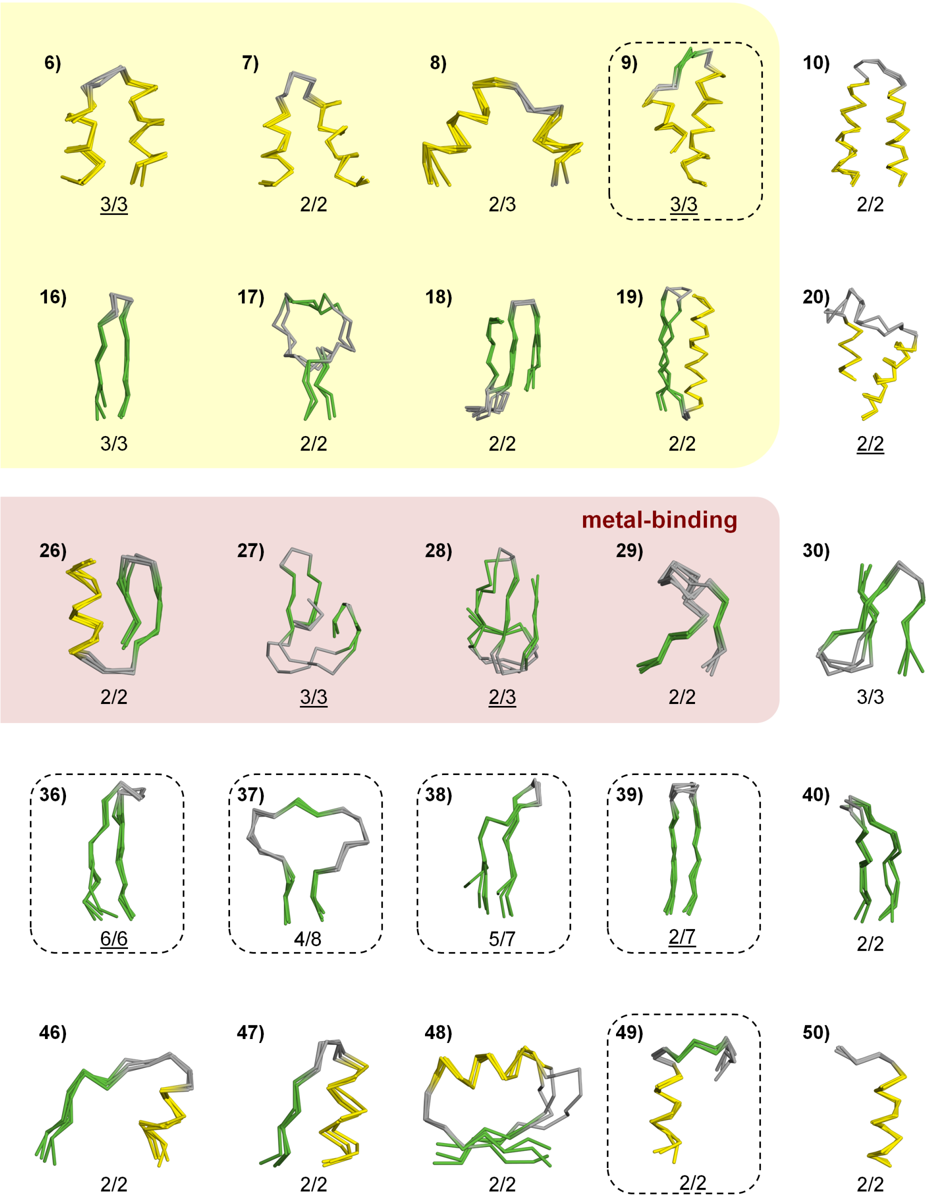


Figure 3.5: Dictionary of ancient peptides (continued). About a third of these fragments (17) have been individually reported before (indicated by an underline), confirming the validity of our approach. The numbers underneath the fragments indicate the number of different SCOP folds and superfamilies they occur in. Nucleic acid-binding, nucleotide-binding, and metal-binding motifs are highlighted in yellow, blue, and red, respectively. Fragments that gave rise to different domains either by accretion or by amplification are indicated by a dotted box.

Second, we also find a high incidence of metal-binding motifs - seven motifs in total - in our set (Fig. 3.5, fragments highlighted with a light red background). In the RNA-based pre-biotic world, RNA was the central macromolecule and carried out most functions, including information storage and catalytic reactions, but there were some reactions, such as redox reactions involving free radicals, that were out of its reach. Since peptides are good chelators of small molecules and can help execute redox reactions via their side chains, it is thought that they were recruited in the RNA world to increase the functional capabilities of RNA [Soeding and Lupas, 2003]. From this perspective, the high incidence of metal-binding motifs in our set might represent an remnant of this evolutionary process. In fact, one of these seven motifs, the FeS-binding peptide (fragment 24), is contained in the ferredoxin-like fold (d.58), which is thought to be one of the five most ancient folds [Caetano-Anolls et al., 2007] .

Finally, of the nine most ancient and basal folds, predicted based on the comparative analysis of proteomes from diverse branches of life, seven (c.37, a.4, c.2, d.58, c.55, b.40, and c.66) contain superfamilies that encompass at least one of our 50 putative ancient peptides. This provides strong proof for the emergence of these fragments in the RNA-peptide world. We note that for the two folds, the TIM $(\beta\text{-}\alpha)_8$-barrel fold (c.1) and the flavodoxin-like fold (c.23), that do not contain any of the fragments from our set a case for a homologous origin has been made [Hoecker et al., 2002, 2004; Bharat et al., 2008]. Furthermore, it is also believed the TIM barrel fold itself arose by amplification of an ancestral module [Soeding, 2005]. Of these nine folds, the P-loop-containing nucleoside triphosphate hydrolase fold (c.37) is thought to be the most ancient one and is involved in binding mononucleotides (e.g. ATP), which are ubiquitous to all known organisms and play a key role in metabolic processes. The DNA/RNA-binding three-helical bundle fold (a.4) and the ferredoxin-like fold (d.58) have been discussed in the earlier paragraphs. The NAD(P)-binding Rossmann fold (c.2) and the S-adenosyl-L-methionine-dependent methyltransferases fold are homologous and encompass a dinucleotide-binding β-α-β motif (fragment 31), which is also found in eight other folds. While the ribonuclease H-like fold (c.55) contains the α-α-β motif (fragment 43), superfamilies belonging to the OB fold (b.40) encompass the helix-hairpin-helix motif (fragment 2) and the zn-ribbon-motif (Fragment 28).

When we initiated this study, we believed that assembly from non-identical fragments may have been one of the primary forces in the evolution of domains, and we expected to find numerous instances of domains built by the recombination of non-identical fragments. However, to our surprise, we did not find even a single domain that contained two or more different fragments from our set, indicating that assembly from non-identical elements may not have been a powerful factor in the evolution of domains. Most fragments occur by themselves in the groups of folds that encompass

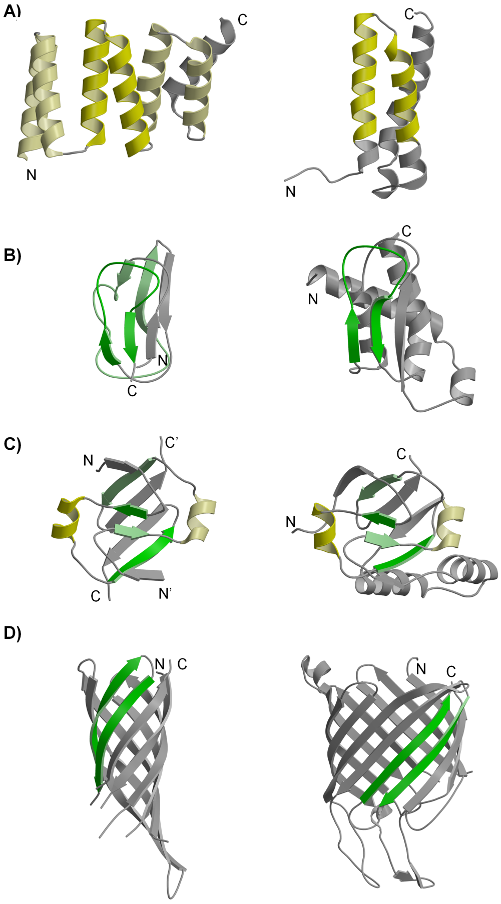


Figure 3.6: Evolution by accretion and amplification. Of the 50 fragments in our set, 14 are found in groups of folds that contain either one copy or multiple copies of the same fragment, or in groups of folds that contain variable number of copies of a common fragment. In all structures α-helices are shown in yellow and β-strands in green. A) The TPR element (fragment 31). Domains of the TRP-like superfamily (a.118.8; 1ELW, shown on the left side) contain multiple homologous copies of the TPR element, while most other folds that encompass this element just have one copy of it (1A17, right). B) The β-β-β-hammerhead motif (fragment 37). The barrel-sandwich hybrid fold is composed of two homologous copies of this motif (1BDO, left), whereas the remaining three folds with this motif include just one copy each (e.g. the α/β-hammerhead fold (d.41); 1QPO, right). C) The transcription factor AbrB (1YFB, left) is a homodimer and contains one copy of the β- $alpha$-β motif (fragment 41) per subunit. MraZ has internal sequence symmetry and contains two homologous copies of the β-α-β motif (1N0E, right). D) Outer membrane β-barrels comprise 4-12 homologies copies of a β-hairpin element (fragment 39); examples include the eight-stranded OmpA (1QJP, left) and the twelve-stranded Tsx (1TLY, right).

them, where they frequently form an integral part of their structures and are generally decorated by additional nonhomologous structural elements. It thus seems plausible to assume that groups of folds containing a shared fragment, decorated in different ways, are a result of piecemeal growth. We also find evidence to suggest that repetition was an important factor in the emergence of domains. We identified many groups of folds that contain either one copy or multiple copies of the same fragment, or of folds that comprise varying number of copies of a homologous fragment. In fact, 14 of the 50 fragments we detected are found in such groups of folds (Fig. 3.5, fragments highlighted by a dotted box). The TPR element (fragment 31), for instance, is found in multiple copies in domains belonging to the TPR-like superfamily (a.118.8), but it is found in a single copy in most other folds that contain it (Fig. 3.6A); the TPR element is found in 8 folds comprising a total of 16 superfamilies. In these other folds, it is decorated by additional structural elements. This suggests that an ancient module corresponding to an α-hairpin was amplified as well as decorated to give rise to the different TPR-element-containing folds. Another example of a fragment that gave rise to different folds by amplification and accretion is the β-β-β motif (fragment 37), which resembles a hammerhead and is found in the barrel-sandwich hybrid fold (b.84), the α/β-hammerhead fold (d.41), and in two other folds (e.29 and f.46). While the barrel-sandwich hybrid fold is pseudo-symmetric and includes two homologous copies of this motif, the remaining three folds, including the α/β-hammerhead fold, contain one copy of it (Fig. 3.6B). We also have fragments in our set that never occur in a single copy within one fold, but do occur in variable number of copies within different folds. Examples of this include the outer membrane β-barrels, which contain between four and twelve homologous copies of a β-hairpin (fragment 39; Fig. 3.6D), and the β-propellers, which are composed of between four and ten homologous copies of a four-stranded β-meander (fragment 35). Finally, we also find instances of folds that are built from multiple homologous copies of the same fragment, either as monomers with internal sequence symmetry or as homo-oligomers with one or more copies per subunit. This, as has also been discussed by previous studies, provides a strong body of evidence for evolution of domains by duplication [Soeding and Lupas, 2003]. The cradle-loop barrels, for instance, either have internal sequence symmetry and contain two homologous copies of the β-α-β motif (fragment 41) or are homodimers with one copy per subunit (Fig. 3.6C). In sum, our findings indicate that repetition and accretion were the key factors in the emergence of domains.

One obvious question that arises here is: how many fragments are we missing? As noted earlier, we found all previously reported sequence- and structure-similar fragments, suggesting that our coverage should be quite exhaustive. However, it is very likely that many ancient fragments are presently represented in just one modern domain superfamily, thereby making their detection impossible. We also anticipate that some cases of ancient fragments remain undetected in our analysis if they involve domains with few homologs of known sequence and structure, as it is not possible to build a reasonable profile HMM in these cases and as structural data is crucial in our analysis for validating the relationships detected using sequence data. With the progress of sequencing and structural genomics projects, this problem should gradually decrease, and might result in the identification of new structural contexts for the fragments identified in our study and may also lead to the detection of some more ancient fragments.

3.3.5 Conclusions

In this section, we have retraced the evolutionary factors that may have shaped the first domains, based on the hypothesis that they evolved by fusion and recombination from an ancient set of peptides, the antecedent domain segments. We compared domains representative of all known fold types using a sequence- and structure-based approach, and identified 50 subdomain-sized fragments that occur in domains of different folds, but exhibit significant similarities in sequence and structure. These fragments are widespread in populous folds, providing compelling evidence for our hypothesis that folded domains arose from simpler fragments. A large number of our fragments are involved in the most ancient functions and are also found in the most ancient folds, indicating that they may represent the detectable remains of the RNA-peptide world from which the first folded domains arose. Finally, our results reveal that repetition and accretion, and not assembly from non-identical fragments, were the main factors in the emergence of domains.

3.3.6 Materials and methods

As noted earlier, the aim of this study was to identify fragments that occur in domains of different folds, but are still similar in sequence and structure. To obtain domains representative of all known fold types, we used the SCOP database (version 1.75) [Andreeva et al., 2008] filtered to a maximum of 20% sequence identity (SCOP20), as provided by the ASTRAL compendium [Brenner et al., 2000]. The SCOP20 dataset contained 7002 domains in total. Multiple alignments were built for each of the SCOP20 domains using the buildali.pl script (with default parameters) from the HHsearch (1.6.0) package. This script uses PSI-BLAST [Altschul et al., 1997] and contains heuristics to reduce the inclusion of nonhomologous sequence segments at the ends of PSI-BLAST sequence matches, the leading cause of high-scoring false positive matches in PSI-BLAST. Profile HMMs were calculated from the alignments using hhmake, also from the HHsearch (1.6.0) package.

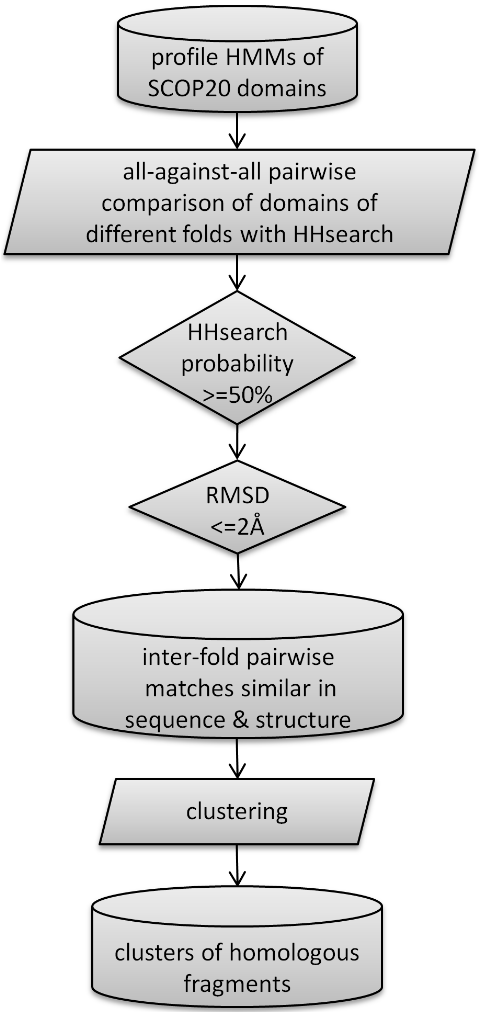

profile HMMs of SCOP20 domains
all-against-all pairwise comparison of domains of different folds with HHsearch
HHsearch probability >=50%
RMSD <=2Å
inter-fold pairwise matches similar in sequence & structure
clustering
clusters of homologous fragments

Figure 3.7: A sequence- and structure-based approach for identifying homologous fragments that reoccur in domains of different folds. We made all-against-all pairwise profile HMM comparisons of domains belonging to different folds in the SCOP20 set using HHsearch. We then removed all pairwise matches with a HHsearch probability of less than 50%, to obtain a list of pairwise matches that show significant similarities in sequence. The structural alignment program TMalign was then used to filter out all pairwise hits with a RMSD of greater than 2 Å. This resulted in a list of pairwise matches between domains of different folds that were based on the presence of sequence- and structure-similar segments. A clustering procedure was then applied to extract groups of domains that shared a common fragment.

We used a structure- and sequence-based approach (Fig. 3.7) to detect homologous fragments that reoccur in different structural contexts. First, we compared all domains of different folds with each other using HHsearch.

We used default settings, but switched off secondary structure scoring (option -ssm 0). This was done to make sure that the obtained matches are not scored superficially high because of the similarity of their predicted secondary structure. All pairwise matches between domains of different folds with a HHsearch probability of less than 50% were ignored. In previous studies, we have noticed that at probabilities of greater than 50% the error rate of HHsearch is quite low. In the next step, we removed all reciprocal hits (of the form *domain A→domain B*, *domain B→domain A*) to reduce redundancy; the hit with the higher probability was retained. In the end, we were left with pairwise matches between domains of different folds that show significant similarities in sequence; these pairwise matches were pooled together for subsequent analysis.

Next, for every pairwise sequence match in our set, we extracted the corresponding segments in structure and aligned them with the structural alignment program TMalign [Zhang and Skolnick, 2005]; all matches with a root-mean-square deviation (RMSD) of less than 2 Åwere removed. We then filtered out all matches in which the aligned segment was shorter than eight amino acids. This was done to remove fragments comprising just a single secondary structure element. Such fragments may have arisen multiple times during the course of protein evolution and as a result, deducing homology between them is fraught with problems. In the end, we obtained a list of pairwise matches between domains of different folds that were similar both in sequence and structure.

Next, we applied a clustering procedure to extract groups of domains that share a homologous fragment. For every pairwise match A→B, where A and B represent distinct regions of two different domains, we iterated through the list of all matches, and pooled together matches of the form: i) A→C (or C→A), where the ends of A were allowed to vary by +/- two residues, ii) A→B' (or B'→A), where the ends of A were allowed to vary by +/- two residues and it was required that the match was to a different segment of domain B (denoted as B'), iii) B→D (or D→B), where the ends of B were allowed to vary by +/- two residues, and iv) B→A' (or A'→B), where the ends of B were allowed to vary by +/- two residues and it was required that the match was to a different segment of A (denoted as A'). This process was repeated with every pooled match until no new matches were found. After applying the clustering procedure, we obtained multiple clusters, each comprising domains of different folds whose similarity was hinged on the presence of a structure- and sequence-similar segment. The obtained clusters were analyzed on a case-by-case basis and the detected relationships were evaluated further by making additional comparisons with domains from a larger set, which comprised domains from the SCOP database filtered to a maximum of 40% sequence identity (SCOP40). We generated sequence and structure alignments for the identified fragments, and assigned fragment boundaries by manual inspection.

Chapter 4

Histones arose from an ancestral helix-strand-helix motif

4.1 Motivation

In the previous chapter, we explored the evolutionary nature of the protein fold space and showed that there is a surprising degree of connectedness in it. We found many incidences of homologous connections between different superfamilies of a fold, indicating that folds may not have had as many independent origins as hitherto assumed. Moreover, we also identified homologous connections between domains of different folds, which arise due to the presence of recurrent sequence- and structure-similar fragments. In a systematic comparison of domains of known fold types, we identified about 50 such fragments and propose that they may be descendants of an ancestral pool of peptide modules from which the first folded proteins arose. In this chapter, we describe in detail one of the fragments we identified, the helix-strand-helix (HSH) motif (Chapter 3, Fig. 3.5-9), which occurs in domains belonging to three different folds, including the histone fold. Histones organize the genomic DNA of eukaryotes into chromatin. The four core histone subunits consist of two consecutive helix-strand-helix motifs and are interleaved into heterodimers with a unique fold. We propose that an antecedent domain segment, corresponding to one helix-strand-helix motif, gave rise divergently to the N-terminal substrate recognition domain of Clp/Hsp100 proteins and to the helical part of the extended ATPase domain found in AAA+ proteins. The histone fold arose subsequently from the latter through a three-dimensional domain-swapping event; this has been evidenced experimentally by a recent study [Hadjithomas and Moudrianakis, 2011].

Parts of this chapter have been adapted with permission from [Alva et al., 2007].

4.2 Background on histones

4.2.1 Eukaryotic histones

Deoxyribonucleic acid (DNA) is the hereditary material in almost all known living organisms (with the exception of RNA viruses). It contains instructions for making proteins and other regulatory elements (e.g. non-coding RNA molecules). Given this central importance of DNA in life, it is not surprising that eukaryotes have evolved a specialized, membrane-enclosed organelle, the nucleus, for storing their DNA. The nucleus contains multiple chromosomes, each encompassing a linear molecule of DNA, that together make up the genome. However, the task of storing DNA in the nucleus is made challenging by the fact that the length of DNA is several times larger than the diameter of the nucleus in most eukaryotes. For example, the total length of human DNA is approximately 2m, but this DNA needs to be packed into a nucleus with a diameter of only 5-10μm. Moreover, it also needs to be packed in a way such that it can be easily accessed for reading various instructions off it.

Eukaryotes achieve the compaction and reversible packaging of their DNA by organizing it into chromatin. The basic structural unit of chromatin is the nucleosome [Kornberg and Thomas, 1974], which consists of 146 base pairs of double-stranded DNA wrapped around an octameric histone core complex [Luger et al., 1997]. This complex is composed of two copies of each of the histone proteins H2A, H2B, H3, and H4, organized as a central (H3-H4)$_2$ tetramer flanked by two H2A-H2B dimers [Arents et al., 1991](Fig. 4.1A). The four core histones are highly conserved across all eukaryotes, from basal eukaryotes (e.g. amoeba) to higher mammals (Fig. 4.2). However, they show very low sequence similarity to each other. Despite this, all core histone subunits share a common fold termed the histone fold; they are composed of three helices separated by two short strap loops (α1-loop1-α2-loop2-α3) and assemble into heterodimers by interleaving the helices into the handshake motif and juxtaposing the strap loops into short parallel β-bridges [Arents et al., 1991]. It has been previously proposed that this fold may have arisen through the duplication of a primordial helix-strand-helix motif [Arents and Moudrianakis, 1993, 1995]. The histone fold is not exclusive to nucleosome core histones; it has also been found in many eukaryotic non-histone proteins mostly involved in DNA metabolism, such as the TATA box binding protein-associated factors (TAFs) and the CCAAT-specific transcription factor CBF [Baxevanis et al., 1995]. It thus appears to be a widespread protein-DNA and protein-protein interaction module.

The core histones have additional N- and/or C-terminal extensions, which are important for gene regulation and chromatin assembly [Luger and Richmond, 1998]. While the part of histone sequences adopting the histone fold is invariable across eukaryotes, the N-terminal tails are more variable. The

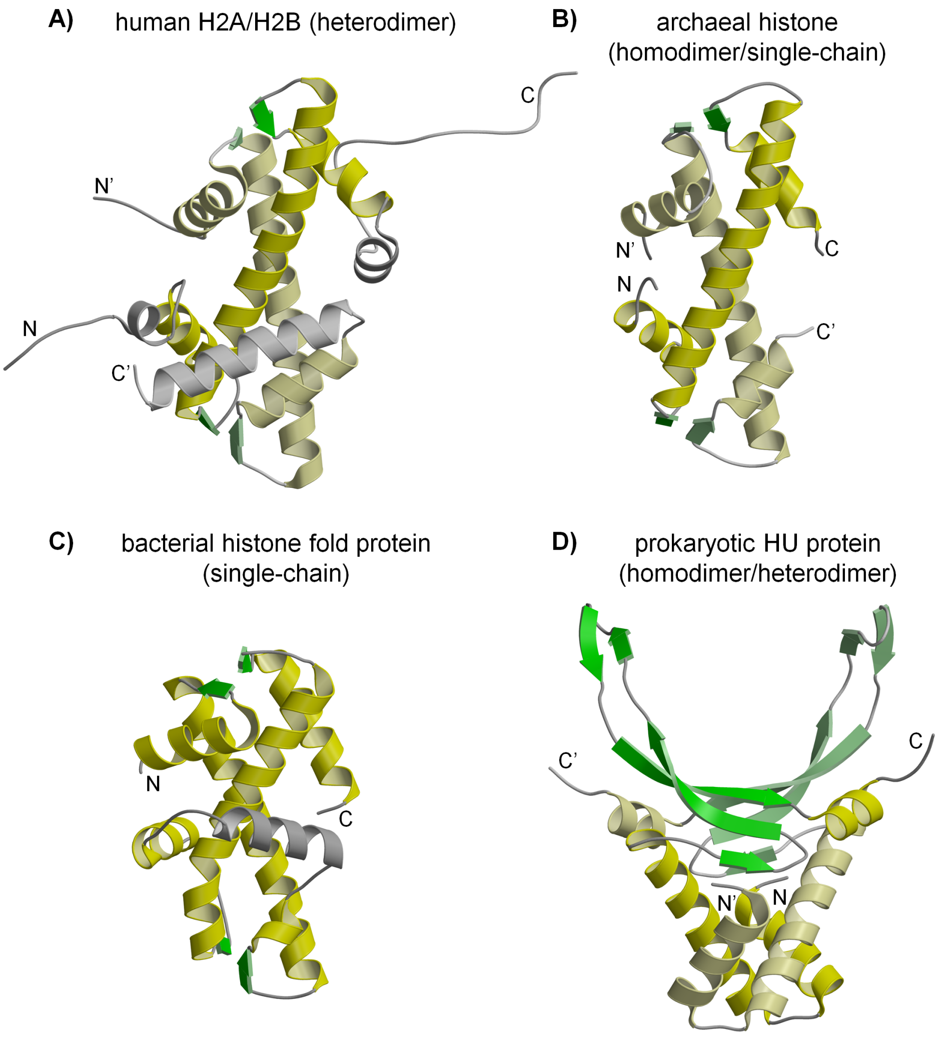


Figure 4.1: Gallery of histones. In all structures, α-helices are shown in yellow, β-strands in green, and loops in gray. Segments not part of the histone fold are also shown in gray. In dimeric structures, monomers are distinguished by light and dark colors. The structures shown are : A) human nucleosome core histones H2A and H2B (PDB 2CV5), B) the histone of archaeon *Methanothermus fervidus* (1B67), C) bacterial histone fold protein from *Thermus thermophilus* (1WWS), and D) HU protein from the cyanobacterium *Anabaena* PCC7120 (1P71).

N-terminal tails of H2A and H2B are hyper variable, whereas the tails of H3 and H4 are more conserved (the multiple alignment of H3 histones in Fig. 4.2 shows the conservation of their tails). The N-terminal tails are found in all eukaryotic lineages, including basal eukaryotes (e.g *Giardia*, *Dictyostelium*, *Entamoeba)*, suggesting that the last common ancestor of all

```
Homo sapiens             MARTKQT--ARKST-GGKAPRKQLATKVARKSAP---ATGGVKKPHRYRP
Mus musculus             MARTKQT--ARKST-GGKAPRKQLATKAARKSAP---ATGGVKKPHRYRP
Tetraodon nigroviridis   MARTKQT--ARKST-GGKAPRKQLATKAARKSAP---ATGGVKKPHRYRP
Arabidopsis thaliana     MARTKQT--ARKST-GGKAPRKQLATKAARKSAP---TTGGVKKPHRYRP
Saccharomyces cerevisiae MARTKQT--ARKST-GGKAPRKQLASKAARKSAP---STGGVKKPHRYKP
Tetrahymena thermophila  MARTKQT--ARKST-GAKAPRKQLASKAARKSAP---ATGGIKKPHRFRP
Paramecium caudatum      MARTKQT--ARKSTAGNKKPTKHLATKAARKTAPAVGAAGGLKKPHKFRP
Dictyostelium discoideum MARTKQT--ARKST-GAKVPRKHIGSKQAHKQTPVSSSSGGVKKVHRFRP
Giardia intestinalis     MARTKHT--ARKTTSATKAPRKTIARKAARKTAS---STSGIKKTGRKKQ
Entamoeba histolytica    MARTKGH--IERPSNKSAKAVKNVAFKAAKKMLS----KDSTKKK-RAHP
Trichoplax adhaerens     MGRTTQSGGLPATSRRKSTPVKRVPASATSSKAN----ESQTRRQKRFKP
                         *.**.       .:     . * :  . : .        .  ::  : :

                                                    α1           β1
                                             hhhhhhhhhhhhhh  h   ee  hhhhhh
Homo sapiens             GTVALREIRRYQKSTELLIRKLPFQRLMREIAQD--FKTDLRFQSSAVMA
Mus musculus             GTVALREIRRYQKSTELLIRKLPFQRLVREIAQD--FKTDLRFQSSAVMA
Tetraodon nigroviridis   GTVALREIRRYQKSTELLIRKLPFQRLVREIAQD--FKTDLRFQSSAVMA
Arabidopsis thaliana     GTVALREIRKYQKSTELLIRKLPFQRLVREIAQD--FKTDLRFQSHAVLA
Saccharomyces cerevisiae GTVALREIRRFQKSTELLIRKLPFQRLVREIAQD--FKTDLRFQSSAIGA
Tetrahymena thermophila  GTVALREIRKYQKSTDLLIRKLPFQRLVRDIAHE--FKAELRFQSSAVLA
Paramecium caudatum      GTVALREIRKYQKSTELLIRKLPFQRLVREIAHE--FQKELRFQSSAVLA
Dictyostelium discoideum GTVALREIRKYQKSTDLLIRKLPFQRLVREIAQE--FKTDLRFQSAAIGA
Giardia intestinalis     GMVAVKEIKKYQKSTDLLIRKLPFSKLVRDIVTSGLSKSDIRFQGAAVEA
Entamoeba histolytica    GAVALTEIKVLQRSTELLLRKAPFQALVREIAQV--SKSDLRFQSAAISA
Trichoplax adhaerens     GTRALMEIRQYQKNTNLLIRKLPFSRVVREVAYD-ITSQHFYWQVEALLA
                         *  *: **:  *:.*:**:** **. ::*::.     . .: :*  *: *

                                  α2              β2       α3
                         hhhhhhhhhhhhhhhhhhhhhh   ee  hhhhhhhhhhh
Homo sapiens             LQEACESYLVGLFEDTNLCVIHAKRVTIMPKDIQLARRIRGERA------
Mus musculus             LQEACEAYLVGLFEDTNLCAIHAKRVTIMPKDIQLARRIRGERA------
Tetraodon nigroviridis   LQEASEAYLVGLFEDTNLCAIHAKRVTIMPKDIQLARRIRGERA------
Arabidopsis thaliana     LQEAAEAYLVGLFEDTNLCAIHAKRVTIMPKDIQLARRIRGERA------
Saccharomyces cerevisiae LQESVEAYLVSLFEDTNLAAIHAKRVTIQKKDIKLARRLRGE--------
Tetrahymena thermophila  LQEAAEAYLVGLFEDTNLCAIHARRVTIMTKDMQLARRIRGERF------
Paramecium caudatum      LQEAAEAYLVGLFEDTNLCAIHARRVTIMSRDIQLARRIRGERF------
Dictyostelium discoideum LQEASEAYLVGLFEDTNLCAIHAKRVTIMPKDIHLARRIRGERS------
Giardia intestinalis     LQESAENYIISLFVDTQLCAEHAKRVTIMKPDMELATRI-GKRIEPEYRK
Entamoeba histolytica    LQEAAEAYLVGLFEDTNLCAIHAKRITIMPKDMQLARRIRGERT------
Trichoplax adhaerens     IQEAAEAFLVSLFEDTNLCAIHAKRVTIMPRDVQLARRIRGRNDILNGK-
                         :**: * :::.** **:*.. **:*:**   *:.** *: *.
```

Figure 4.2: Multiple sequence alignment of nucleosome core histone H3 from diverse eukaryotes ranging from protists to mammals. Secondary structure elements corresponding to the histone fold are indicated. The degree of conservation of residues in each alignment column is indicated as follows: (*) identical residues in all sequences, (:) highly conserved column, and (.) weakly conserved column.

eukaryotes already possessed the N-terminal tails.

In addition to the four core histones, eukaryotes possess the linker histone H1, which does not feature the histone fold. Higher eukaryotes contain multiple variants of the linker histone (e.g. humans have 11 variants) and these variants have been implicated to have specific functions [Izzo et al., 2008]. The linker histones connect nucleosomes to form higher order structures and thus are integral parts of the chromatin structure. They are also involved in other important cellular processes [Izzo et al., 2008]. Like the core histones, the linker histones are also found in all eukaryotes; however, in comparison they are more divergent. The linker histones of multicellular eukaryotes show a tripartite architecture comprising a central globular domain and two flanking, largely unstructured, amino- and carboxyl-terminal domains. The globular domain adopts a winged-helix fold comprising a three-helix bundle followed by a β-hairpin [Zarbock et al., 1986; Ramakrishnan et al., 1993]. In contrast to multicellular eukaryotes, only some protists contain this globular domain.

4.2.2 Prokaryotic histones

The situation in prokaryotes is quite different to the one in eukaryotes. While essentially all eukaryotes use two copies of each of the four core histones to compact their DNA, prokaryotes contain a variety of nucleoid-associated proteins (NAPs) that are distinct from histones. Like eukaryotic histones, most prokaryotic NAPs bind DNA in a sequence-independent way and have the ability to bend, wrap, or bridge DNA.

Alongside with many NAPs such as Lrp, Alba, and HU, many archaea, including all methanogens, most euryarchaeotes, and some crenarchaeotes, also possess nucleosome-like structures made up of histones [Sandman et al., 1990; Pereira et al., 1997; Sandman and Reeve, 2001; White and Bell, 2002; Cubonov et al., 2005; Sandman and Reeve, 2006]. These histones form homo- and heterodimers (Fig. 4.1B) and assemble into tetramers, which may reflect an ancestral form of the central part of the eukaryotic nucleosome octamer, the (H3-H4)$_2$ tetramer [Bailey et al., 1999]. Most archaeal histones are essentially the histone fold, with no N-terminal or C-terminal extensions. However, there are archaeal histone variants that either have C-terminal extensions [Li et al., 2000] or additional domains. Archaeal histone subunits are occasionally duplicated on a single polypeptide chain [Fahrner et al., 2001], a form observed in eukaryotes only in the histone-like domain of the *son of sevenless* protein [Baxevanis et al., 1995]. Unlike eukaryotic core histone sequences that show high conservation, archaeal histone sequences exhibit a degree of variation.

Bacteria do not have eukaryote-like nucleosomes, but they have nucleoid proteins with histone-like properties [Drlica and Rouviere-Yaniv, 1987; Lynch et al., 2003; Guo and Adhya, 2007]. These proteins do not contain the histone fold. One particular NAP called HU has properties similar to histones and has been dubbed as the prokaryotic counterpart of histones. It is very abundant in *E. coli* and is thought to be important for the compaction of its genome [Rouvire-Yaniv et al., 1979; Kano et al., 1991]. HU exists as homodimers in almost all bacteria except *enterobacteriaciae*, where it exists as heterodimers of related subunits. These subunits contain a α-hairpin with a long protruding arm comprising five β-strands (Fig. 4.1D).

Recently, structures of two homologs of archaeal single-chain histones were reported from the bacteria *Aquifex aeolicus* (1R4V) [Qiu et al., 2006] and *Thermus thermophilus* HB8 (1WWS; Fig. 4.1C). Further homologs appear in the genomes of a few, phylogenetically diverse bacteria. It thus seems likely that the histone fold originated in the common ancestor of eukaryotes and archaea, and spread into some bacteria through lateral gene transfer.

4.3 Remote homologs of histones

We first found the remote homologs of histones described in this chapter in an all-against-all application of HHpred [Soeding, 2005] to the SCOP database. Subsequently, we systematically looked for homologs of the histone fold by searching the SCOP25 database [Andreeva et al., 2004] with sequences from the three protein families with this fold: archaeal histones (SCOP a.22.1.2), nucleosome core histones (a.22.1.1), and TBP-associated factors (TAFs; a.22.1.3). Results of the searches are presented in Fig. 4.3. As expected, proteins from these three families identified each other as their best matches with high probability; most of the hits were found at probability values of 70%-100% (Fig. 4.3).

Surprisingly, their subsequent matches were consistently to the helical part of the extended ATPase domains found in AAA+ proteins (the C-domain; Fig. 4.4B; c.37.1.20) [Ammelburg et al., 2006; Neuwald et al., 1999], also at high probability values. AAA+ proteins are a group of ATPases that mediate ATP-dependent unfolding and disaggregation of macromolecules [Ammelburg et al., 2006]. They are composed of one or more copies of an extended P-loop ATPase domain, followed by an α-helical subdomain, referred to as the C-domain.

Indeed, we also found significant matches to yet a third protein family, the N-terminal domain of Clp/Hsp100 proteins (Clp N-domain; Fig. 4.4C; a.174.1.1) [Zeth et al., 2002]. Clp/HSP100 proteins are molecular chaperones that promote the ATP-dependent degradation of proteins by forming a complex with ClpP protease [Zeth et al., 2002].

Reciprocal searches with a set of C-domain sequences confirmed the similarity of these protein families (Fig. 4.3). While C-domains identified each other as their best matches, they also made matches to histones and N-domains at high probabilities.

4.4 Analysis of sequence and structure conservation

The surprising aspect of these findings is that histones, C-domains, and Clp N-domains belong to three different folds (Fig. 4.4A-C). Histones are

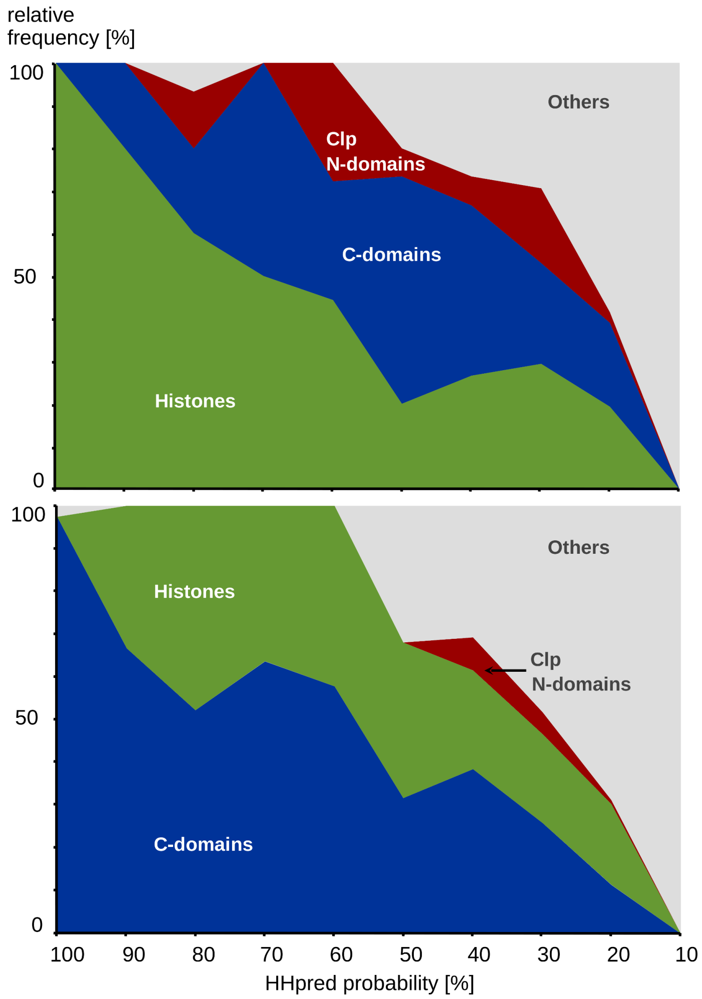


Figure 4.3: Results of HHpred searches of the SCOP25 database with histone sequences and C-domains (reproduced from [Alva et al., 2007]). The relative frequencies of SCOP families encountered in the searches are plotted against the HHpred probabilities as described in the *Materials and methods* section (4.8). Searches with histones are shown in the top panel and searches with C-domains in the bottom panel. Histones (SCOP a.22) are colored in green, C-domains (SCOP c.37.1.20; also includes the misclassified a.49.1.1) in blue, Clp N-domains (SCOP a.174) in red, and others in gray.

dimeric, interleaved helical bundles (Fig. 4.4A), as described in the *Background* section (4.2.1). C-domains are four-helix bundles composed of two consecutive helix-strand-helix motifs (Fig. 4.4B) [Ammelburg et al., 2006]. Clp N-domains, finally, are multi-helical domains formed by the repetition of a four-helical motif (Fig. 4.4C) [Zeth et al., 2002]. Although these three protein families have different topologies, they all incorporate two copies of

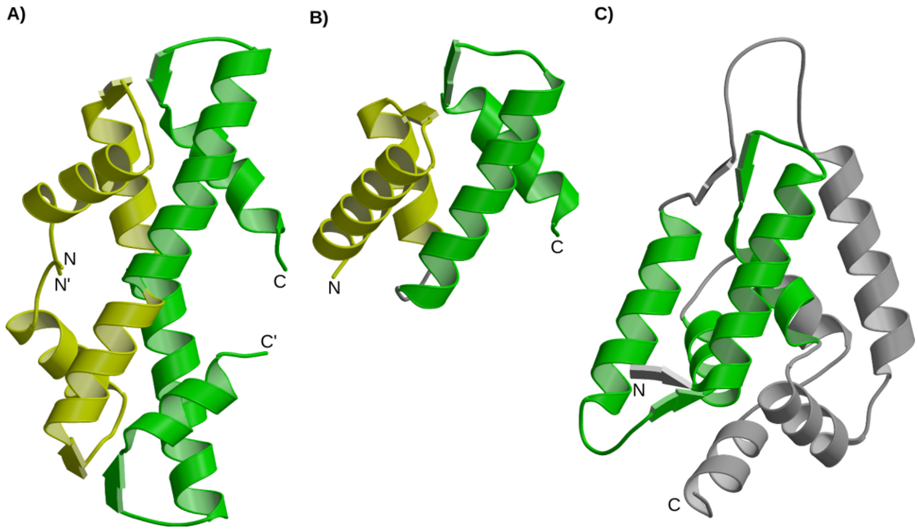


Figure 4.4: The structure of histones, C-domains, and Clp N-domains (reproduced from [Alva et al., 2007]). A) The histone of *Methanothermus fervidus* (1B67); the N-terminal helix-strand-helix motif in each subunit is colored yellow and the C-terminal motif green. B) The C-domain of the helicase RuvB (1IN4); the motifs are colored as in the histone subunits. C) Clp N-domain of ClpA (1K6K); the two helix-strand-helix motifs are colored green.

the helix-strand-helix motif, which engages in the formation of a short parallel β-bridge. In the histone dimer, the β-bridge is formed by the association of one helix-strand-helix motif from each monomer, in the C-domain by the association of the two motifs consecutive in the polypeptide chain.

The similarities detected by HMM-to-HMM comparison are limited to these helix-strand-helix motifs. Histones and C-domains both contain two consecutive copies of the motif and can be aligned over essentially their entire length (Fig. 4.5). Clp N-domains contain two motifs decorated by two helices and each motif has its best matches to the C-terminal motif of histones and C-domains (Fig. 4.5). The sequence alignment shows extensive similarity in the hydrophobic patterns of the three folds, but no highly conserved residues other than two alanines in the core of the second helix-strand-helix motif, which allow for close packing interactions at the crossover point between the helices.

A structural comparison of the three folds shows that C-domains can be superimposed onto one half of the histone fold with root-mean-square deviations (RMSD) of around 1.5Å (Table 4.1 and Fig. 4.6). The main difference between the two folds lies in the fact that the two helix-strand-helix motifs of C-domains are connected by a hinge region, while they are continuous in histones, requiring dimerization to form the hydrophobic core (Fig. 4.6). The structural alignment between histones and Clp N-domains

```
Histone fold           α1             β1    α2----------------α2                     β2       α3
1b67_A (2)   -gelpIAPIGRIIKNAGAe------RVS-DDARIALAKVL----------EEMGEEIASEAVKLAKHAGR-KTIK---AEDIELARKMf
1taf_A (19)  --pkdAQVIMSILKELNVq------EYE-PRVVNQLLEFT----------FRYVTSILDDAKVYANHARK-KTID---LDDVRLATEVt
1taf_B (3)   gssisAESMKVIAESIGVg------SLS-DDAAKELAEDV----------SIKLKRIVQDAAKFMNHAKR-QKLS---VRDIDMSLKV-
1tzy_A (21)  glqfpVGRVHRLLRKGNYae-----RVG-AGAPVYLAAVL----------EYLTAEILELAGNAARDNKK-TRII---PRHLQLAIRNd
1tzy_B (33)  rkesySIYVYKVLKQVhpdt-----GIS-SKAMGIMNSFV----------NDIFERIAGEASRLAHYNKR-STIT---SREIQTAVRLl
1tzy_C (62)  rklpfQRLVREIAQDFKtdl-----RFQ-SSAVMALQEAS----------EAYLVGLFEDTNLCAIHAKR-VTIM---PKDIQLARRIr
1tzy_D (27)  -qgitKPAIRRLARRGGvk------RIS-GLIYEETRGVL----------KVFLENVIRDAVTYTEHAKR-KTVT---AMDVVYALKRq
C-domain of AAA+       α1             β1      α2     hinge            α3             β2       α4
1in4_A (181) tVKELKEIIKRAASLMDV-------EIE-DAAAEMIAKR-srgt------PRIAIRLTKRVRDMLTVVKA-DRIN---TDIVLKTMEVl
1r6b_A (350) sIEETVQIINGLKPKYeahhdv---RYT-AKAVRAAVEL-avkyindrhlPDKAIDVIDEAGARARLMpvskrkktvnVADIESVVARi
1lv7_A (326) dVRGREQILKVHM--rrv-------PLApdidAAIIARG-tpgfs-----GADLANLVNEAALFAARGNK-RVVS---MVEFEKAKDKi
1ny5_A (314) rKEDIIPLANHFLKKFsrkyakevegFT-KSAQELLLSYpwygn------VRELKNVIERAVLFSegk----FID---RGELSCLV---
1g8p_A (220) dVETRVEVIRRRDTYD|V-------EAP-NTALYDCAAL-cialgs--dgLRGELTLLRSARALAALEGA-TAVG---RDHLKRVAT--
Clp-N motif                          [β2]                             α1             β1       α2
1k6k_A (4)   ------------------------[qpt]----------------------QELELSLNMAFARAREHRH-EFMT---VEHLLLALLSn
1k6k_A (82)  ------------------------[mln]----------------------LSFQRVLQRAVFHVQSSGR-NEVT---GANVLVAIFSe
1khy_A (8)   ------------------------[qps]----------------------NKFQLALADAQSLALGHDN-QFIE---PLHLMSALLNq
1khy_A (85)  ------------------------[rlt]----------------------QDLVRVLNLCDKLAQKRGD-NFIS---SELFVLAALEs
```

Figure 4.5: Sequence comparisons of histones, C-domains, and Clp N-domains (reproduced from [Alva et al., 2007]). Multiple sequence alignment of representative members of each fold. Residues in helices are colored yellow in histones, cyan in C-domains, and green in Clp N-domains; residues in β-bridges are colored red. Structurally equivalent residues are shown in capital letters and residues forming the hydrophobic core are shown in bold. The sequences are labeled by their PDB codes; the numbers in brackets refer to the residue number for the first residue in the alignment.

Table 4.1: Data for the structural alignment in Fig. 4.6

PDB ID	Fold	Aligned length	RMSD [Å]
1B67(A:5-66, B:105-166)	Histone	124	0.00
1TAF(A:21-84, B:9-70)	Histone	123	1.23
1TZY(C:66-131, D:30-92)	Histone	124	1.60
1IN4(A:182-250)	C-domain	64	1.83
1LV7(A:327-396)	C-domain	63	2.15
1K6K(A:4-38)	N-domain	32	1.52
1K6K(A:81-118)	N-domain	33	1.27

is also in the range of 1.5Å RMSD, but extends only over the C-terminal helix-strand-helix motif of histones.

4.5 Evolutionary implications

4.5.1 The histone fold evolved from the C-domain through a 3D domain-swapping event

The structural and sequence similarity between histones and the C-domains of AAA+ proteins, despite differences in their overall topology, suggests that they are evolutionarily connected. We propose 3D domain-swapping as the mechanism that accounts for their structural differences. 3D domain-swapping is a process by which two or more identical proteins exchange a domain to form interlocked oligomers [Bennett et al., 1995], in which all of the packing interactions that stabilize the monomer are present. The swapped portions can range from a single secondary structure element to an entire domain. In the simplest case the native fold, normally constituted by a single 'closed' monomer, is reconstituted by two so-called 'open' monomers. This reciprocal swap leads to a homodimer, whereas the runaway domain swap, in which swapping propagates along an axis in an open-ended manner, has been proposed to contribute to amyloid fibril formation [Janowski et al., 2001; Sambashivan et al., 2005; Guo and Eisenberg, 2006].

Up to now, about 40 proteins have been shown to be able to undergo 3D domain-swapping [Liu and Eisenberg, 2002], and several studies indicate a physiological role of this mechanism in allostery and signal transduction [Piccoli et al., 1988; Gotte et al., 1999; Schymkowitz et al., 2001]. A precondition is the presence of a flexible loop or hinge, about which the swapped elements can rotate in order to form a pair of 'open' monomers. The primary intervention by which 3D domain swaps have been engineered into monomeric proteins is through the shortening of the hinge, thus preventing the packing of part of the protein into its native location and forcing a swap, such as in domain 1 of lymphocyte antigen CD2 [Murray et al., 1995], staphylococcal

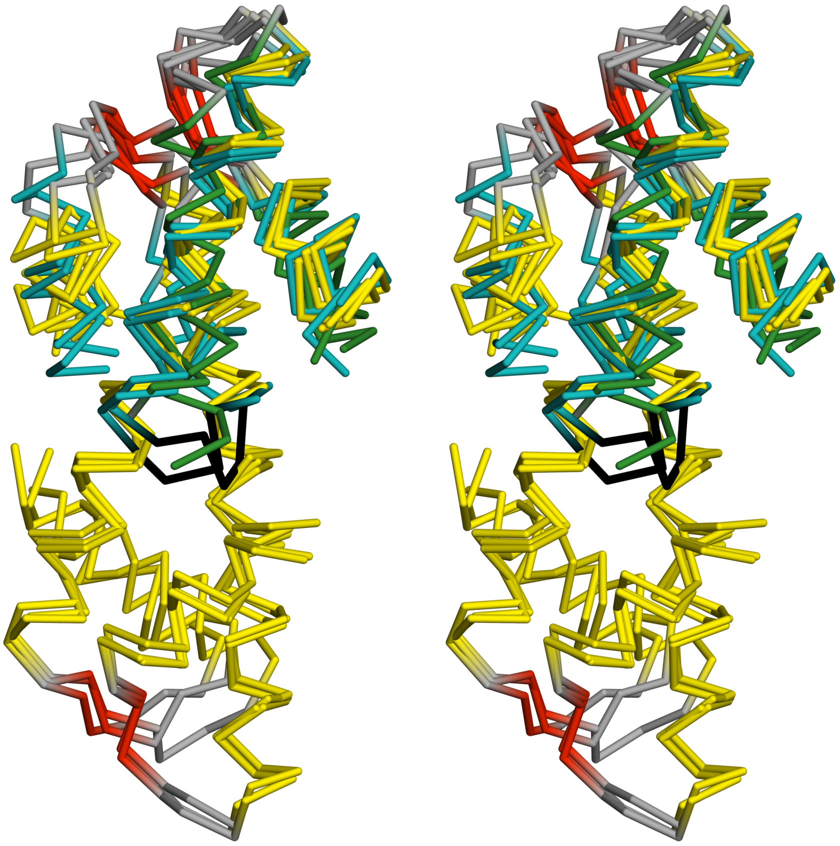


Figure 4.6: Structure comparisons of histones, C-domains, and Clp N-domains. The superposition was made using the archaeal histone HMfA (1B67) as a reference structure. Quantitative information on the results of the superposition is provided in Table 4.1. Residues in α-helices are colored yellow in histones, cyan in C-domains, and green in Clp N-domains. Residues in β-bridges are colored red and the hinge region of C-domains is highlighted in black.

nuclease [Green et al., 1995], single-chain Fv fragments [Kortt et al., 1994; Perisic et al., 1994], and in a three-helix bundle designed by *Ogihara et al.* [2001].

Our results suggest that such a shortening of the hinge region, which connects the two helix-strand-helix motifs of the AAA+ C-domain, led to a 3D domain swap. The event caused head-to-tail dimerization of monomers, which thereby recovered the lost interactions between the two helix-strand-helix motifs, and resulted in the emergence of the histone fold (Fig. 4.7). Following the proposal that domain-swapping might contribute to protein evolution [Bennett et al., 1995; Kinch and Grishin, 2002], we present here the first concrete example.

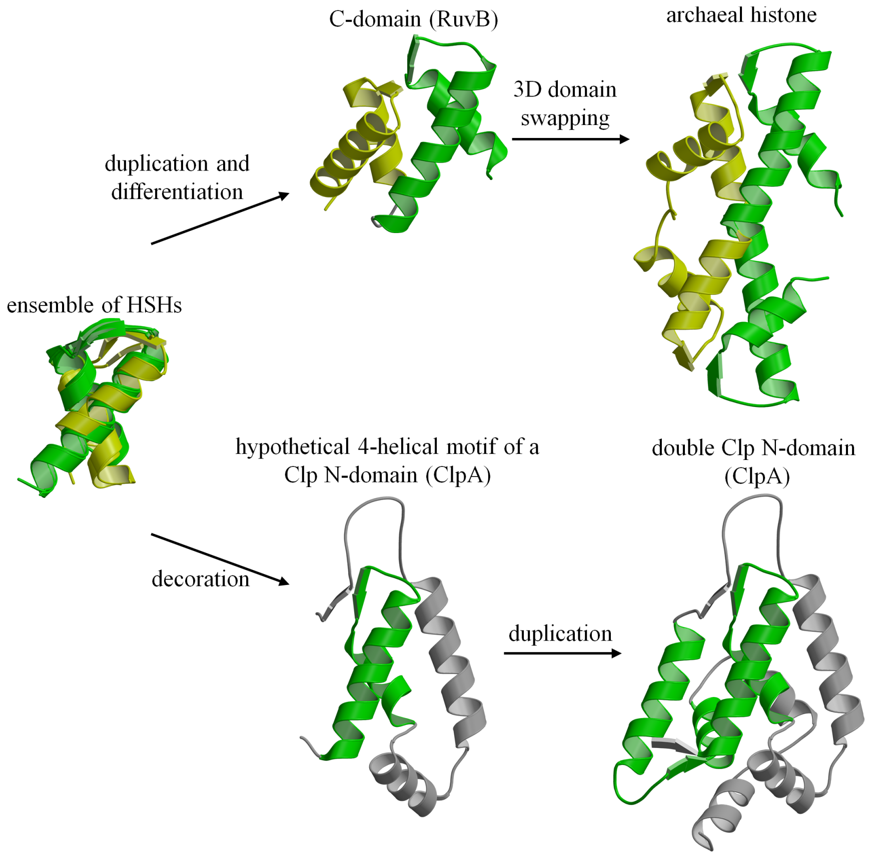

C-domain (RuvB)
archaeal histone
3D domain swapping
duplication and differentiation
ensemble of HSHs
hypothetical 4-helical motif of a Clp N-domain (ClpA)
double Clp N-domain (ClpA)
decoration
duplication

Figure 4.7: Evolutionary scenario for the origin of three folds from an ancestral helix-strand-helix motif (reproduced from [Alva et al., 2007]). The coloring and representative proteins are as in Fig. 4.4. The superimposed ensemble of helix-strand-helix motifs consists of motifs from the following proteins: yellow (1IN4: residues 181-212; 1B67: 4-33), green (1IN4: 216-251; 1B67: 134-166; 1K6K: 82-115). A primordial helix-strand-helix motif, first, gave rise divergently to the Clp N-domain and the AAA+ C-domain through two independent events of duplication and fusion. The histone fold arose subsequently from the C-domain by 3D domain-swapping event.

4.5.2 A primordial helix-strand-helix motif

The helix-strand-helix motif (Chapter 3, Fig. 3.5-9), which is at the core of the similarity between histones and C-domains, is also found in Clp N-domains, which assume yet a third fold. Here, the motif is decorated with two C-terminal helices, and two copies of this extended, four-helical motif are fused in antiparallel orientation. Thus, three different folds appear to have been built from a common helix-strand-helix motif. The helix-strand-helix motif, we propose, is an ancestral peptide, which gave rise divergently

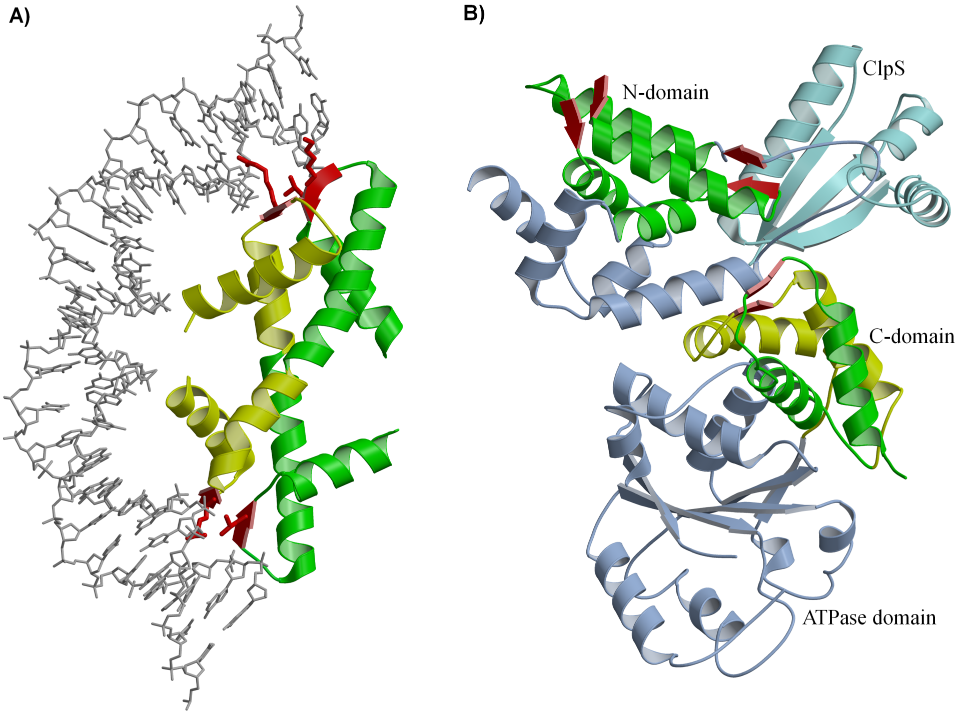


Figure 4.8: Involvement of the β-bridge in macromolecular interactions (reproduced from [Alva et al., 2007]). The coloring of helix-strand-helix motifs is as in Fig. 4.4, except the β-bridges are colored red. A) The histone H3-H4 complex bound to DNA (1S32); residues of the β-bridges engaged in interactions with the phosphate backbone are shown in stick representation. B) ClpS in complex with ClpA (1LZW, 1R6B); the ATPase domain is in light blue and ClpS in cyan.

to the Clp N-domain and the AAA+ C-domain through two independent events of duplication and fusion (Fig. 4.7). The C-domain then evolved into the histone fold by 3D domain-swapping.

4.6 Functional implications

An interesting structural feature common to all three folds is the presence of one or two short, parallel β-bridges formed by the strands of the helix-strand-helix motifs. In histones, these β-bridges provide the main site of interaction with the phosphate backbone of DNA (Fig. 4.8A). In Clp N-domains, one of the two β-bridges binds the adaptor molecule ClpS [Zeth et al., 2002; Maurizi and Xia, 2004] (Fig. 4.8B). Although the binding sites of the AAA+ C-domains have not been characterized yet, it thus seems attractive to propose that here also the single β-bridge formed in this domain represents the main binding site. C-domains play an important role in sensing the nucleotide bound by the AAA+ proteins [Botos et al., 2004; Ogura et al., 2004; Diemand and Lupas, 2006] and are located close to the substrate-binding N-domains (Fig. 4.8B), projecting radially at the circumference of the hexameric ring complex. We note in this context that C-domains are frequently rich in positively charged residues and that in the Lon protease, the C-domain has been implicated in interactions with DNA [Lee et al., 2004]. We propose that the helix-strand-helix motif served as a scaffold for the formation of parallel β-bridges. Ancestrally, these bridges bound proteins, but in a few C-domains they also acquired the ability to bind DNA, eventually leading to histones as proteins that only bind DNA at these sites.

4.7 Recent developments

4.7.1 Experimental evidence for the role of domain-swapping in the origin of histones

Recently, *Hadjithomas et al.* [2011] experimentally tested our hypothesis on the emergence of the histone fold. They investigated whether inserting a C-domain-like hinge loop into the middle of the central helix of the histone fold (Fig. 4.1B) would yield a four-helix-bundle as seen in the C-domains. They compared many C-domains to each other and to histones, and identified a Gly-Thr-Pro (GTP) motif shared by the hinge regions of the *E. coli* proteins RuvB and FtsH for their engineering experiment. Because glycine and proline are helix breakers, and threonine has a small side-chain, the GTP motif was deemed to be an appropriate hinge for this experiment. The GTP motif was inserted in the middle of the central helix, between residues Glu33 and Glu34, of the homodimeric archaeal histone HMfB from *Methanothermus fervidus.* The wild-type and the engineered protein were analyzed using sedimentation equilibrium in the analytical ultracentrifuge, and interestingly, while the wild-type is a dimer as expected, the engineered protein is a soluble and stable monomer with properties similar to C-domains. The authors suggest that the insertion of the GTP motif disrupts the central α-helix and thereby also disturbs the handshake motif, allowing the histone fold to collapse on itself to yield the C-domain fold. This provides compelling evidence for our proposal that the histone fold evolved from the C-domain via a domain-swapping event.

4.7.2 Discovery of the C-domain outside the AAA+ superfamily

The C-domain remained exclusive to AAA+ proteins until its recent discovery in a second superfamily of proteins, the DNA polymerase epsilon (ε) subunit B superfamily [Mkiniemi et al., 1999]. This superfamily contains B subunits of eukaryotic family B polymerases α, δ, and ε. *Nuutinen et al.* [2008] recently solved the structure of the N-terminal domain of human DNA polymerase ε (Dpoe2NT) and it was found to resemble the C-domain of AAA+ proteins in structure. We note that HHpred searches find C-domains as their best matches and thus indicating that Dpoe2NT is homologous to C-domains. Dpoe2NT is specific to the B subunit of eukaryotic DNA polymerases ε and is not found in the B subunits of other family B polymerases [Nuutinen et al., 2008]. In addition to the Dpoe2NT domain, B subunit of eukaryotic DNA polymerases contain two more domains, an OB-fold domain and a calcineurin-like phopshoesterase domain.

4.8 Materials and methods

We obtained histone and Clp N-domain sequences from the ASTRAL compendium [Brenner et al., 2000] as defined by the SCOP (version 1.71) [Andreeva et al., 2004] folds a.22 and a.174, respectively. We reduced this set to less than 25% pairwise identity at 90% length coverage using BLASTCLUST [Altschul et al., 1990]. C-domains are not characterized as a separate fold in SCOP; we extracted their sequences from the extended AAA-ATPase family (c.37.1.20) of the SCOP database by a procedure described by *Ammelburg et al.* [2006] and also reduced this set to less than 25% pairwise identity. Our final dataset comprised 16 histones, 18 C-domains, and two Clp N-domains.

We used these sequences to search the SCOP25 database for homologs with HHpred [Soeding, 2005], at default parameters and a probability cutoff of 10%. The SCOP25 database is a version of SCOP filtered for a maximum of 25% pairwise sequence identity. For each group, we pooled all search results and tabulated the frequencies at which various SCOP families appeared at each probability, binned at 10% intervals.

The histone, C-domain, and Clp N-domain structures were superimposed interactively in Swiss-PDB viewer [Guex and Peitsch, 1997]. We chose the archaeal histone HmfA (1B67) as the reference structure, as it made the highest number of connections both in sequence and structure searches. Quantitative information for the superimposition is listed in Table 4.1. The complex shown in Fig. 4.8B, consisting of ClpS, N-domain, and the first AAA+ domain of ClpA, was generated by superimposing the N-domains of the structures 1R6B (N-domain and the AAA+ domains) and 1LZW (N-domain in complex with ClpS) from *E. coli.*

Chapter 5

The GD box: a recurrent non-contiguous structural motif

5.1 Motivation

Thus far, we have concentrated on homologous repeats that are found multiple times within a domain and on sequence- and structure-similar fragments that reoccur in domains of different folds. We favor the hypothesis that these local similarities observed in modern domains represent the remnants of the ancient set of short peptides from which the first folded domains emerged. In addition to containing homologous fragments, unrelated domains frequently also show recurrent local substructures with completely different amino acid sequences. These elements comprise two or more secondary structure elements in specific geometric arrangements [Rao and Rossmann, 1973], termed supersecondary structures, some of which are widespread in proteins, e.g. α-hairpins, β-hairpins, and β-α-β motifs. In fact, more than 50% of the residues in the most highly populated folds are found in these three elements [Salem et al., 1999]. Although these supersecondary structural elements show considerable similarity in structure, they exhibit totally different sequences, suggesting that they are a result of some general principles governing the structural organization of proteins rather than common ancestry. Indeed, because there are only a limited number of energetically favorable ways to pack secondary structural elements [Finkelstein and Ptitsyn, 1987], unrelated proteins tend to converge upon similar local structures. In contrast, sequence space is essentially infinite and many sequences are compatible with a particular local structure.

All supersecondary structures described so far in literature are formed

Parts of this chapter have been adapted with permission from [Alva et al., 2009].

by adjacent segments in the polypeptide chain. However, given the central role of non-local interactions in the formation of protein structure (see for example *Minor et al.* [1996)], it seems reasonable to assume that some widespread supersecondary structures should be formed by non-contiguous secondary structure elements. In this chapter, we describe a widespread non-contiguous fragment, termed the GD box, that is present in both homologous and analogous contexts. We initially found this fragment in a group of topologically distinct but homologous β-barrels - the cradle-loop barrels [Coles et al., 1999, 2005, 2006; Ammelburg et al., 2007], which we describe in more detail in Chapter 6. The GD box is similar both in sequence and structure and comprises two short unpaired β-strands connected by an orthogonal type-II β-turn and a non-contiguous β-strand forming hydrogen bonds with the β-turn. Using structure-based analysis, we have detected 518 instances of the GD box in a non-redundant subset of the SCOP database comprising 3771 domains. Apart from the cradle-loop barrels, this motif is also found in a diverse set of non-homologous folds including other topologically related β-barrels. Since non-local interactions are fundamental in the formation of protein structure, systematic identification and characterization of other non-contiguous supersecondary structural elements is likely to prove valuable to protein structure modeling, validation, and prediction.

5.2 Background on the GD box

Cradle-loop barrels are a group of topologically distinct but homologous β-barrels, whose similarity is based on the presence of a recurrent antecedent domain segment corresponding to a β-α-β-motif (Chapter 3, Fig. 3.5-41). These barrels illustrate the evolution of folded proteins from simple oligomers of one fragment to the emergence of complex catalysis [Coles et al., 1999, 2005, 2006; Ammelburg et al., 2007]. In order to capture the relationship between such distinct folds originating from the same basic supersecondary structure, we proposed a new protein classification level, the metafold [Alva et al., 2008] (a detailed account on this is provided in Chapter 6). The cradle-loop metafold comprises four distinct folds, the double-psi barrel (Fig. 5.1A), the swapped-hairpin barrel, the RIFT barrel, and the C-terminal barrel of bacterial fluorinating enzyme, each built from two copies of the conserved β-α-β-element, either as monomers with internal sequence symmetry or as homo-dimers with one copy per subunit. A characteristic feature of the β-α-β-element is a conserved 11 amino acid sequence motif, [h]-x-[h]-x(2)-G-[p]-x-[h]-x-[h], where [h] is hydrophobic and [p] polar. As the polar residue is frequently aspartate, we named this motif the GD box. In all cradle-loop barrels, this motif is structurally highly conserved and comprises two short unpaired β-strands connected by an orthogonal diverging type-II β-turn (Fig. 5.1C). Diverging β-turns were first described in the I-sites library,

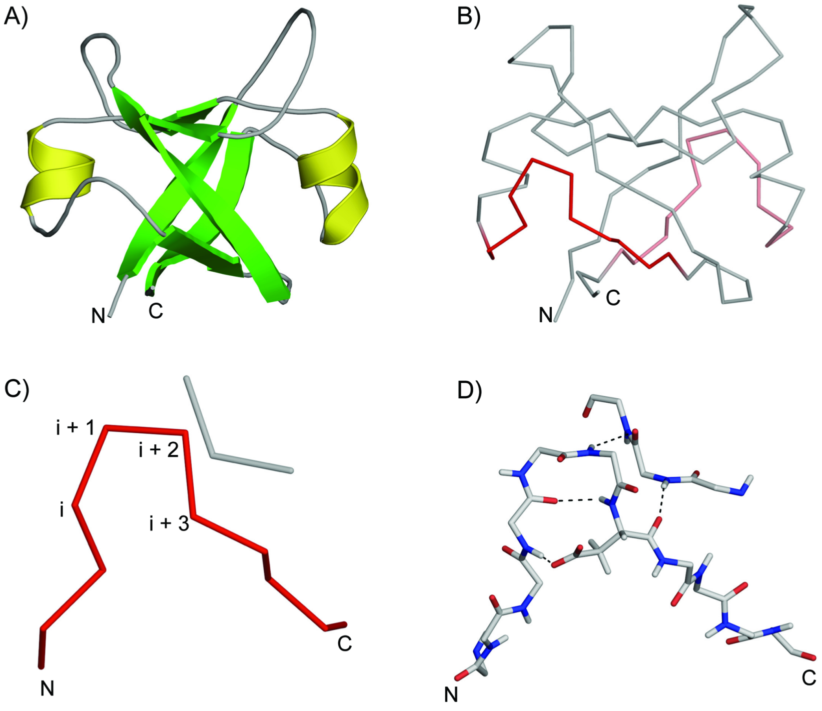


Figure 5.1: The GD box element (reproduced from [Alva et al., 2009]). A) The double-psi barrel fold of VatN-N (PDB 1CZ4, residues 1-91) is shown in cartoon representation. α-helices are colored in yellow and β-strands in green. B) The double-psi barrel fold of VatN-N is shown in backbone representation. The unpaired hairpins (residues 34-44 and 77-87) from the two GD box elements are shown in red. C) The first GD box element from VatN-N (residues 34-44, 53-55) is shown. The positions corresponding to the type-II β-turn are marked. D) Detailed view of the hydrogen-bonding network of the first GD box element.

which contains motifs that correlate both in sequence and structure [Bystroff and Baker, 1998]. The glycine occupies the i+2 position of the β-turn and the side chain of the polar residue in i+3 accepts a hydrogen bond from the residue at position i. The backbone of the residues in i+2 and i+3 forms further hydrogen bonds to a non-contiguous β-strand from the symmetry-related half of the barrel (Fig. 5.1D). Thus the GD box element is an example of a non-contiguous supersecondary element; to our knowledge the first identified.

5.3 GD box elements in known structures

We wondered if the GD box is only found in cradle-loop barrels or also in other non-homologous folds. To investigate this we analyzed the occurrence

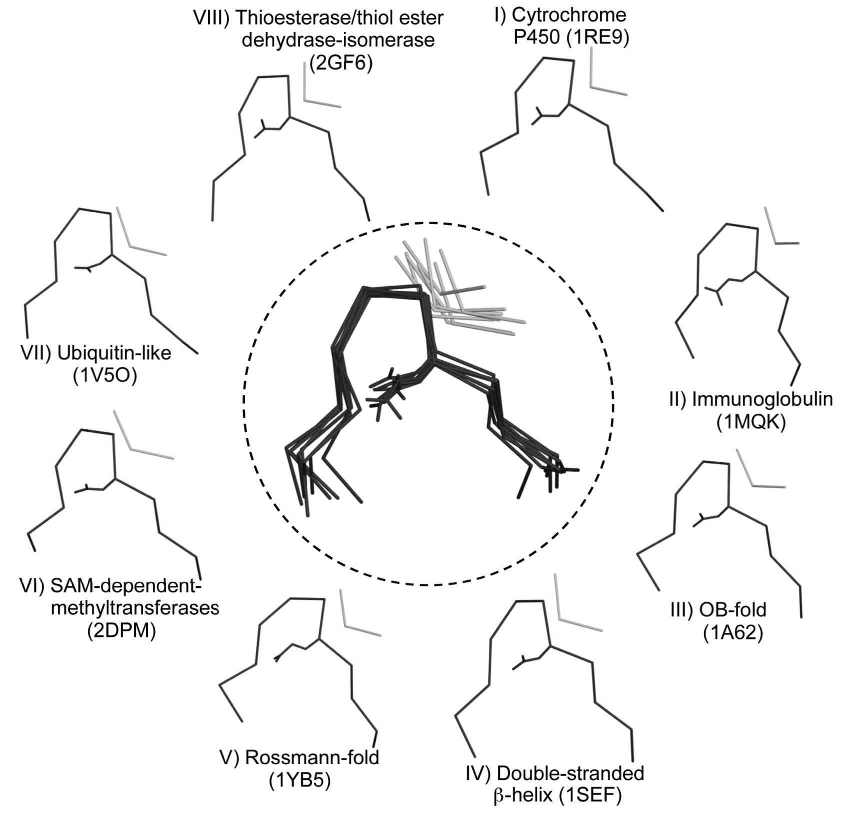

VIII) Thioesterase/thiol ester dehydrase-isomerase (2GF6)
I) Cytochrome P450 (1RE9)
VII) Ubiquitin-like (1V5O)
II) Immunoglobulin (1MQK)
VI) SAM-dependent-methyltransferases (2DPM)
III) OB-fold (1A62)
V) Rossmann-fold (1YB5)
IV) Double-stranded β-helix (1SEF)

Figure 5.2: A Gallery of GD box elements (modified from [Alva et al., 2009]). GD box elements from eight different folds are shown. The unpaired hairpin is shown in black and the non-contiguous segment in gray. A structural superposition of the GD box elements is shown in the center. The structures shown are I) 1RE9: residues 300-310, 290-292, II) 1MQK: chain L, 11-21, 77-79, III) 1A62: 89-99, 54-56, IV) 1SEF: 223-233, 208-210, V) 1YB5: 143-153, 173-175, VI) 2DPM: 183-193, 232-235, VII) 1V5O: 71-81, 9-11, and VIII) 2GF6: 68-78, 11-13.

of the GD box in proteins of known structure. We used structure comparisons to detect GD boxes in SCOP25, a subset of the SCOP database filtered at 25% sequence identity. Since the GD box adopts a very similar structure in all cradle-loop barrels, we chose the first GD box from the N-terminal domain of the archaeal AAA chaperone VAT (PDB 1CZ4, residue 34-44) as the query structure. We were not aiming at detecting all GD box elements in an exhaustive manner, so we used a moderately generous root-mean-square deviation (RMSD) cutoff of 1.5Å. Because one of the goals of this study was to establish whether the GD box forms the same non-contiguous interactions in all its embodiments, we did not consider the presence of hydrogen bonds between the residues at positions i+2 and i+3 of the β-turn and a

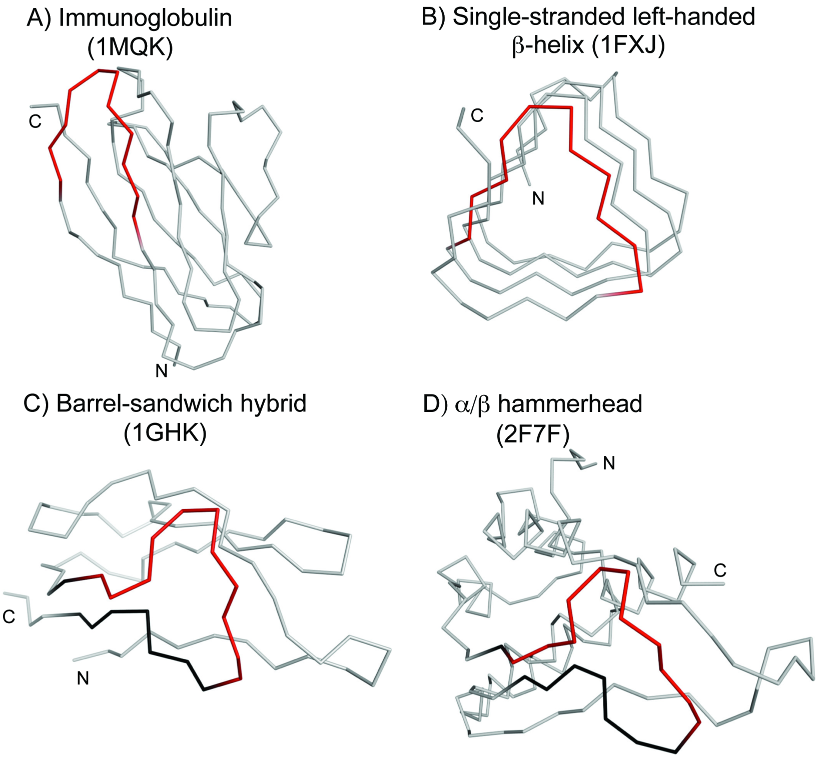


Figure 5.3: A Gallery of GD box-containing folds (reproduced from [Alva et al., 2009]). The contiguous segment of the GD box element is shown in red in all the structures. The β-β-β-motif containing the GD box element is shown in black in C) and D). The structures shown are: A) 1MQK: chain L, B) 1FXJ: chain A, 252-329, C) 1GHK, and D) 2F7F: chain A, 4-140.

non-contiguous β-strand in our searches.

We detected a total of 518 GD boxes in 420 distinct domains, which are classified into 134 folds in SCOP25; some domains contained multiple copies of the element. At 3771 domains total, this means that slightly more than 10% of all domains in SCOP25 contain at least one GD box. In all but 32 of the detected cases, at least one of the residues at position i+2 and i+3 of the β-turn formed a backbone hydrogen bond with a residue in a non-contiguous β-strand. The non-contiguous interaction can thus be considered a general feature of GD boxes.

As a control, we wanted to assess the proportion to which type-II β-turns form GD boxes. We therefore detected all type-II β-turns in SCOP25 using Promotif [Hutchinson and Thornton, 1996] and found 6,327 instances, indicating that less than 10% of all type-II β-turns are part of GD box elements.

Apart from the cradle-loop barrels, the GD box is found in other topo-

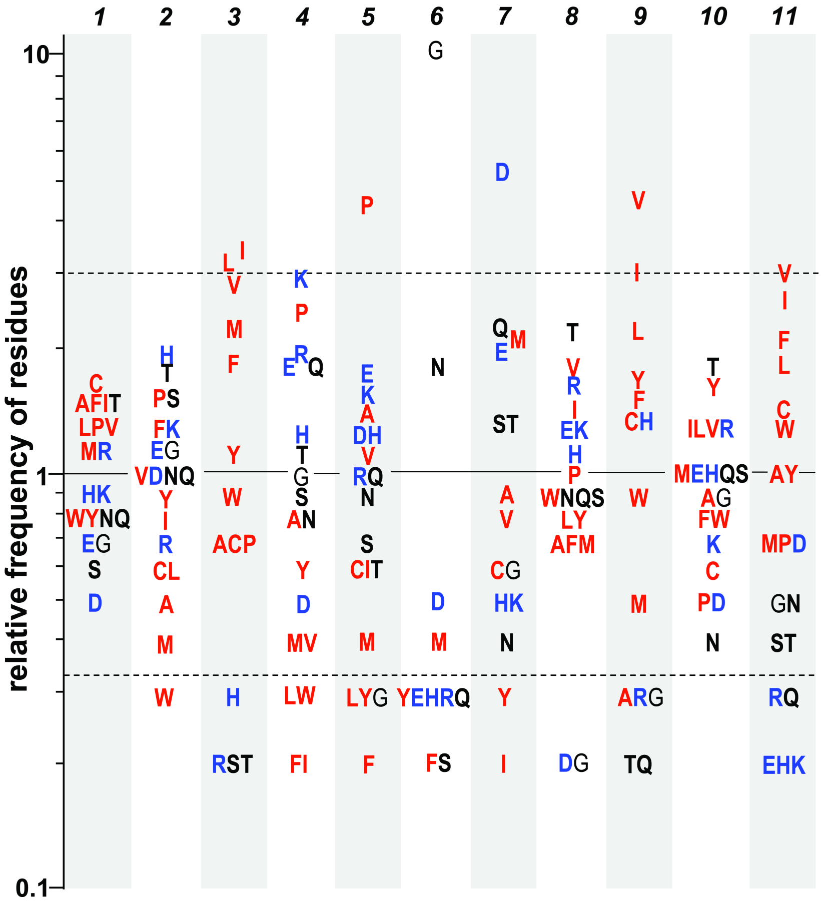


Figure 5.4: The relative preference of occurrence for the 20 amino acids at the 11 positions of the GD box (reproduced from [Alva et al., 2009]). Dotted lines indicate the position of residues that are three times more frequent and three times less frequent than expected, respectively. Hydrophobic residues are colored in red, charged residues in blue, and uncharged hydrophobic residues in bold black. The 11 positions of the GD box are indicated.

logically related, but non-homologous barrels [Alva et al., 2008], as well as in a large number of folds that are not related either evolutionarily or topologically to the cradle-loop barrels (Fig. 5.2). The latter include the OB fold (SCOP: b.40), the immunoglobulin-like fold (b.1; Fig. 5.3A), the NAD(P)-binding Rossmann fold (c.2), and the ubiquitin-like fold (d.15). Some folds contain multiple copies of the GD box element; for example, each half-barrel of the cradle-loop barrels contains its own copy of the GD box element and single-stranded (b.81; Fig. 5.3B) and double-stranded β-helices (b.82) contain multiple overlapping copies. In most cases, the GD box element coordinates a β-strand that is distant in the linear polypeptide sequence and in some cases, this β-strand is contributed by a different subunit al-together. Thus, in all homodimeric cases, such as in the transition state regulator AbrB from *Bacillus subtilis* (PDB 1YFB), the non-contiguous strand originates from the symmetry-related monomer. We propose that the GD box acts as a structural tether, allowing the polypeptide chain to snap into the final folded conformation once the hydrophobic collapse has produced a molten globule and brought the non-contiguous parts into approximate vicinity, and stabilizing the structure once folding is complete. This may explain the widespread and frequently analogous representation of GD boxes in a broad range of β folds.

5.4 Sequence features of the GD box motif

Although the searches for the GD box elements were made by structural comparison, their sequences (considering only the unpaired hairpin) follow the same characteristic pattern as in the cradle-loop barrels: [h]-x-[h]-x(2)-G-[p]-x-[h]-x-[h] (Table 5.1 , Fig. 5.4). Hydrophobic residues are strongly favored at positions 3, 9, and 11, and to lesser extent at position 1. Positions 4, 5, 6, and 7 correspond to the four positions (i, i+1, i+2, and i+3) of the type-II β-turn; at these positions hydrophilic residues are favored. Position 6 is dominated by glycine and to lesser extent by asparagine. Type-II β-turns favor cysteine, serine, and lysine at position i+3 [Hutchinson and Thornton, 1994]; however, in GD box elements the i+3 position is predominantly occupied by aspartic acid and to lesser extent by glutamine and glutamic acid.

5.5 Discussion

5.5.1 Evolutionary consequences

Because structure space is finite, unrelated proteins tend to converge upon similar local structures. This is reflected in the fact that over half of the residues in the most highly populated folds are found in one of the three most common supersecondary structures, i.e. α-α-hairpins, β-β-hairpins, and β-α-β-elements [Salem et al., 1999]. In contrast, sequence space is essentially infinite and many sequences are compatible with a particular local structure. Sequence convergence should thus be highly unlikely. For this reason, statistically significant sequence similarity is considered the best marker for homology. However, short, structurally constrained parts of the polypeptide chain do converge on specific sequence motifs, as described here for the GD box. In such cases, and particularly where these structurally constrained elements form a substantial part of the entire polypeptide chain, it seems possible that the convergent sequence motifs would lead to a level of overall sequence similarity that could be interpreted (erroneously) as indicative of

Table 5.1: Positional propensities for each of the 11 positions of the GD box[a]; (reproduced from [Alva et al., 2009])

Residue	1	2	3	4	5	6	7	8	9	10	11
ILE	1.3	0.8	3.2	0.2	0.6	0	0.2	1.4	3	1.2	2.7
PHE	1.3	1.2	1.7	0.2	0.2	0.2	0.1	0.7	1.4	0.8	2.1
VAL	1.2	1	2.9	0.4	1.1	0	0.8	1.7	4.8	1.2	3
LEU	1.2	0.6	3.1	0.3	0.3	0	0.1	0.8	2.2	1.2	1.7
TRP	0.8	0.3	0.9	0.3	0	0	0	0.9	0.9	0.8	1.2
MET	1.1	0.4	2.1	0.4	0.4	0.4	2	0.7	0.5	1	0.7
ALA	1.3	0.5	0.7	0.8	1.3	0.1	0.9	0.7	0.3	0.9	1
GLY	0.7	1.1	0.1	1	0.3	10.4	0.6	0.2	0.3	0.9	0.5
CYS	1.5	0.6	0.7	0	0.6	0	0.6	0.1	1.3	0.6	1.3
TYR	0.8	0.9	1.1	0.6	0.3	0.3	0.3	0.8	1.5	1.6	1
PRO	1.2	1.4	0.7	2.4	4.5	0	0	1	0	0.5	0.7
THR	1.3	1.6	0.2	1.1	0.6	0.1	1.3	2.2	0.2	1.8	0.4
SER	0.6	1.4	0.2	0.9	0.7	0.2	1.3	0.9	0.1	1	0.4
HIS	0.9	1.7	0.3	1.2	1.2	0.3	0.5	1.1	1.3	1	0.2
GLU	0.7	1.1	0.1	1.7	1.6	0.3	1.9	1.2	0	1	0.2
ASN	0.8	1	0	0.8	0.9	1.7	0.4	0.9	0	0.4	0.5
GLN	0.8	1	0.1	1.7	1	0.3	2.2	0.9	0.2	1	0.3
ASP	0.5	1	0	0.5	1.2	0.5	5.4	0.2	0.1	0.5	0.7
LYS	0.9	1.2	0.1	2.9	1.4	0.1	0.5	1.2	0.1	0.7	0.2
ALA	1.1	0.7	0.2	1.8	1	0.3	0.1	1.6	0.3	1.2	0.3

[a] Amino acids are ranked in order of decreasing hydrophobicity. Positions 4, 5, 6, and 7 correspond to the four positions of the type-II β-turn.

homology.

We therefore wanted to evaluate to what extent sensitive sequence comparison methods, such as those based on the comparison of profile Hidden Markov Models (HMMs), would return scores for GD box-containing proteins that are normally seen between homologous proteins. To this end we made pairwise comparisons of profile HMMs for all GD box-containing domains and clustered them in CLANS (see *Materials and methods* (5.6)). At settings at which we recover the cradle-loop barrels as a cluster, whose homologous origin we have documented in a series of studies, most other GD box-containing folds did not exhibit connections to the cradle-loop barrels or to each other. The convergent similarity of GD boxes is thus not sufficient to suggest homology between evolutionarily unrelated domains, even when these are being compared by methods calibrated for the detection of very distant relationships.

Two other clusters formed in the three-dimensional map. One contained the Rossmann folds, whose connections relied however on a different supersecondary structure, a dinucleotide-binding β-α-β-element whose homologous origin has been discussed previously [Lupas et al., 2001]; our dictionary of ADSs includes this element (Chapter 3, Fig. 3.5-21). The other contained the barrel-sandwich hybrid and the α/β-hammerhead folds. The similarity between these two folds relies on a GD box-containing β-β-β-element, which resembles a hammerhead (Fig. 5.3C, Fig. 5.3D; this element is also present in our dictionary of ADSs (Chapter 3, Fig. 3.5-37). The barrel-sandwich hybrid fold is pseudo-symmetric and contains two homologous copies of the β-β-β-element. The α/β-hammerhead fold contains one copy of the β-β-β-element. We conclude that these two folds most likely have arisen from an ancestral β-β-β-element - the barrel-sandwich hybrid fold by duplication and the α/β-hammerhead fold by accretion.

5.5.2 Application to tertiary structure prediction

Protein folding is still an unsolved problem [Kryshtafovych et al., 2005; Lupas, 2008]. One encouraging approach has been through methods, such as ROSETTA [Simons et al., 1999], which use fragment libraries to predict tertiary structure by assembling local structural features. However, these methods are mainly successful for domains with less than about 100 residues. One reason for their poor scalability may lie in the fact that they do not consider non-local interactions, which become progressively more important with the size of the fold. Enriching these fragment libraries with widespread non-contiguous supersecondary structures that have clear sequence-structure patterns, such as the GD box, should make it possible to include knowledge of non-local interactions into this approach.

A problem with using non-local interactions as restraints is that, while the contiguous part of the element will be recognizable based on its sequence pattern, the non-local interaction partner will be difficult to identify. In NMR structure calculation from ambiguously assigned cross-peaks, as well as in protein-protein docking, one faces similar problems. Indeed, fragment assembly could be viewed as a protein-protein docking task. One can deal with the ambiguities by using ambiguous distance restraints [Nilges, 1995; Dominguez et al., 2003]: the restraint is defined on the whole set of possible distances; the one that is most compatible with the overall structure is then picked automatically during the structure calculation process. Such a restraint-based approach would significantly reduce the computational complexity, and thus, would allow modeling and prediction of larger proteins as well.

5.6 Materials and methods

For this study, we used the SCOP database [Andreeva et al., 2004] (version 1.73) filtered for a maximum of 25% sequence identity. After filtering out all NMR structures and all X-ray structures with a resolution of worse than 2Å, we obtained a subset comprising 3771 domains.

For structure comparisons, we used an implementation of the rigid-body superposition algorithm described by *Challis et al.* [1995]. Only Cα atoms were considered for superposition. The GD box motif from the N-terminal domain of the archaeal AAA chaperone VAT (PDB code 1CZ4, residues 34-44) was compared to all 11 residue fragments from the SCOP25 dataset. Fragments with an RMSD less than 1.5Åwith respect to the probe fragment were pooled together. All fragments without a type-II β-turn were removed from this set. We classified as canonical GD boxes all fragments in which at least one of the residues at position i+2 and i+3 of the type-II β-turn was involved in hydrogen-bonding interactions with a non-neighboring residue; the remaining fragments were classified as non-canonical. The programs Promotif [Hutchinson and Thornton, 1996] and HBPlus [McDonald and Thornton, 1994] were used to detect β-turns and to calculate hydrogen bonds, respectively. The non-canonical fragments were further analyzed in the context of the full protein: if the residues in the β-turn formed hydrogen bonds to a non-neighboring residue, the fragment was also classified as canonical.

The positional propensities for each position of the GD box were calculated as $P_i(a) = \frac{F_i(a)}{F(a)}$, where $P_i(a)$ is the positional propensity of amino acid a at position i, $F_i(a)$ is the frequency of a at position i and $F(a)$ is the background frequency of a in the dataset.

For sequence comparisons, we used HHsearch [Soeding, 2005], which is a highly sensitive homology search method based on the pairwise comparison of hidden Markov models (HMMs). We built multiple sequence alignments for all GD box-containing domains using the buildali.pl script from the HHsearch package. Profile HMMs were calculated from the alignments using hhmake (from the HHsearch package) with default settings. We then performed all possible pairwise comparisons between them using HHsearch with default settings and clustered them by their pairwise P-values using CLANS [Frickey and Lupas, 2004]. Clustering was done to equilibrium in 2D at a P-value cutoff of 1.0e-03 using default settings.

Chapter 6

On the classification of proteins based on natural descent

6.1 Motivation

Proteins are generally composed of one or more autonomously folding modules, termed domains, which retain their overall fold and frequently also their function when they reoccur in other unrelated proteins (see Chapters 1 and 2 for more details). Domains thus represent units of structure, function, and evolution in modern proteins, a knowledge that has been immensely beneficial for the study of proteins. It has, for instance, become an exceedingly common practice in modern molecular biology to delineate a protein-of-interest into its constituent domains ahead of complex experimentation, to obtain first clues of its likely function and overall structure, which frequently forms the basis of subsequent experimental analysis. This approach is often employed for characterizing the three-dimensional structure of large proteins that are hard to work with experimentally. Such large proteins are first divided into their domains, structures of which are then determined individually and assembled to get an approximation of the full-length structure. Owing to these reasons, systems for classifying protein structures have become an essential tool in molecular biology.

In the last two decades, a number of different structural classification systems have been developed, examples include Dali Domain Dictionary [Dietmann et al., 2001], CATH [Greene et al., 2007], and SCOP [Andreeva et al., 2008]. These systems differ in the way they are generated: while the Dali Domain Dictionary is largely automated, relying on the popular structure comparison program DALI, CATH combines automated and manual pro-

Parts of this chapter have been adapted with permission from [Alva et al., 2008].

cesses, and SCOP is based mostly on manual assignments. Despite these differences, all structural classifications systems follow a similar hierarchical scheme to capture the duality between homologous and analogous contributions to the properties of modern proteins. They classify proteins by combining homologous criteria at lower hierarchical levels with analogous criteria at upper levels. SCOP, for instance, groups related domains into families, related families into superfamilies, structurally similar superfamilies into folds, and folds into secondary structure classes. Despite being generally useful, this mode of classification frequently introduces inconsistencies within and between systems, and often also fails at capturing certain distant evolutionary relationships. A case in point is the recent substantiation of several incidences of homologies between domains of different folds, either due to the detection of homologous fold change (e.g. circular permutation) [Grishin, 2001a] or the detection of ancient peptides, as detailed in Chapter 3. However, because current systems classify proteins by using homologous criteria at lower levels and analogous criteria at upper levels, they make the implicit assumption that homologous domains always have the same fold, and thus fail at capturing these inter-fold relationships. To alleviate this issue, we propose to use the metafold [Day et al., 2003; Alva et al., 2008] as the next hierarchical level above the fold, encompassing a group of topologically related folds for which a homologous relationship has been substantiated. We see this as an important step on the way to a classification of proteins by natural descent. In this chapter, we present our ideas on the metafold concept using cradle-loop barrels, which are a group of homologous barrels with topologically distinct but related folds, as an example.

6.2 The metafold concept

Despite showing an appreciable amount of agreement, each of the protein classification systems offers its own view of fold space, with many differences becoming apparent upon detailed, protein-by-protein comparison [Day et al., 2003]. One of the most important source of disagreement between classification systems are differing domain definitions, but in a fair number of cases, disagreements also arise because the same domain is assigned to different folds. In order to understand these problems, it is useful to consider the definition of a fold as a conserved, topologically distinct arrangement of secondary structures in a domain, with extensions and insertions peripheral to the fold treated as decorations. A fold change occurs when one or more secondary structure elements within the fold alter their nature and/or their topology. Clearly a wide latitude in judgment is possible with respect to domain boundaries, to what constitutes a decoration, and to the degree of topological change necessary to separate structures into different folds. To alleviate the discrepancies arising from subjective estimates of fold similarities and differences, *Daggett and coworkers* introduced the metafold concept as a consensus method designed to reveal fold similarities relatively independently of the methods used to compare protein domains [Day et al., 2003]. In essence, domains are considered part of the same metafold if their topological similarity is recognized by multiple classification systems.

Although the metafold concept is clearly helpful in obtaining a more unified view of fold space, we think that its usefulness is limited by the fact that it does not address a fundamental source of tension within and between classification systems, namely the coexistence of homologous and analogous classification criteria. Both SCOP and CATH explicitly use homology for classification, SCOP in the first two hierarchical levels (family and superfamily) and CATH in the first level. Clearly, what one perceives as homologous has a profound influence on drawing domain boundaries and on labeling parts of the structure as decorations to a conserved core. These decisions, which have a substantial arbitrary component, by necessity shape the higher levels of the classifications. Furthermore, by placing homology at the base and topological considerations into the upper layers, both systems make the implicit assumption that homologous proteins always have the same fold. As has become increasingly clear in recent years, this is not the case and multiple events have been described that can lead to fold change in evolution: point mutations, indels, topological substitutions, circular permutations, strand swaps, strand and hairpin invasions, 3D domain swaps, and chimeric fusions [Grishin, 2001a; Kinch and Grishin, 2002; Andreeva and Murzin, 2006; Andreeva et al., 2007]. In such cases, if the evidence for a common ancestry is obvious, homology typically trumps topological dissimilarity, leading to the grouping of homologous proteins into the same fold, even when they have undergone changes that would cause analogous proteins to be classified into separate folds. For example, the C2 domains of synaptogamin I (PDB 1RSY) and phospholipase C (PDB 1QAS), which are related by a circular permutation event, are classified into the same superfamily (b.7.1) in SCOP. However, in cases where evidence for a common ancestry is weak, classification decisions are often subjective. For instance, while β-propellers, which are circular folds with between four and eight repeats of a four-stranded β-meander [Chaudhuri et al., 2008], are classified into five different folds in SCOP (b.66-b.70), outer membrane β-barrels, which are closed barrels with between four and twelve repeats of a β-β hairpin [Remmert et al., 2009], are classified into different superfamilies within one fold (f.4). Furthermore, as detailed in Chapter 3, the boundaries between folds have also been broken due to the detection of homologous fragments in domains of different folds.

In order to address the tensions arising in classification systems from fold changes between homologous proteins and from the detection of homologous fragments in domains of different folds, we think that the metafold would more usefully be viewed as the next hierarchical level above the fold, bringing together groups of topologically similar folds for whom a homologous relationship has been substantiated in at least one case. This core group of member folds could be usefully expanded by a periphery of candidate folds, which are related to the core members by a topological change known to occur between homologous proteins, albeit without homology having been substantiated in that particular case. Proteins known to be analogous to members of the core group should be excluded from the metafold. Thus, topologically similar proteins may end up in different metafolds, based on evidence of their descent. In such cases, these proteins will also need to be assigned to different folds. Although this may seem to subvert the concept of fold, we note that this concept has always had a large arbitrary component and that some larger folds are currently subdivided without easily recognizable reasons. The goal of our proposal is to ultimately eliminate all analogous criteria from protein classifications.

Although more complicated and also fuzzier than the definition of *Daggett and coworkers*, our metafold definition offers several advantages: it addresses the homology-analogy problem directly, it has considerably greater explanatory power with regard to the fold space it maps, it acknowledges explicitly that the classification is more robust and better supported in some cases (the core groups) than in others (the peripheries), and it provides a first step toward a comprehensive classification of proteins by descent.

In the following, we would like to explain our ideas on metafolds using the example of cradle-loop barrels, which we have studied in detail for almost a decade.

6.3 The cradle-loop barrel metafold

6.3.1 The core group

Our interest in the cradle-loop barrels began with the structure determination of VatN, the substrate recognition domain of the AAA protein VAT [Coles et al., 1999]. The N-terminal part of this domain (VatN-N) forms one of the most topologically complex folds known, consisting of duplicated β-α-β (Chapter 3, Fig. 3.5-41) units that are completely interdigitated to form a six-stranded barrel (Fig. 6.1A and Fig. 6.2). The β1-β2 loops of each unit cross over the symmetry-related $\beta 2'$ strand, giving the fold an unusual, knotted appearance and its name: the double-psi β-barrel [Castillo et al., 1999]. In each β-α-β unit, the helix is connected to the last strand by a conspicuous Gly-Asp motif, which we named the GD box (see Chapter 5).

In order to understand how such a complex fold could have evolved, we undertook an extensive bioinformatics and structural study with the aim of identifying possible precursors with simpler topologies. We focussed on several groups of proteins that showed sequence similarity to VatN-N. Of

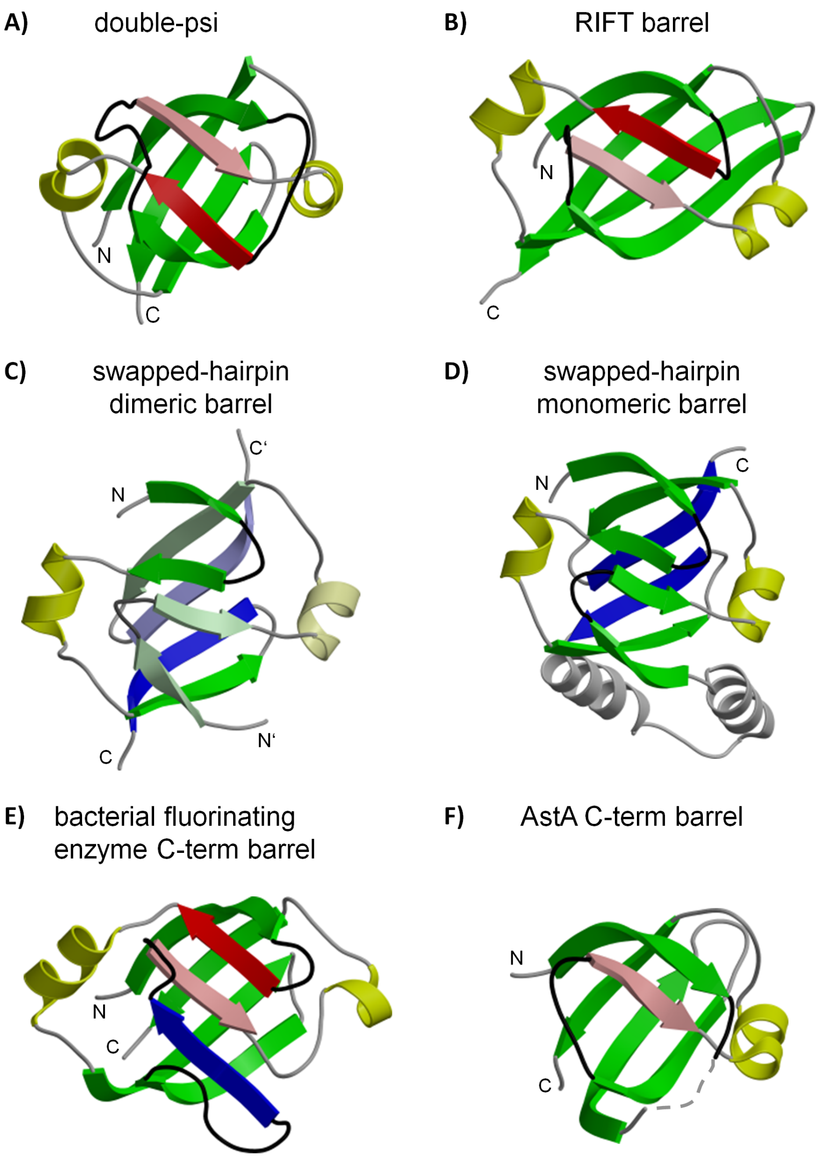
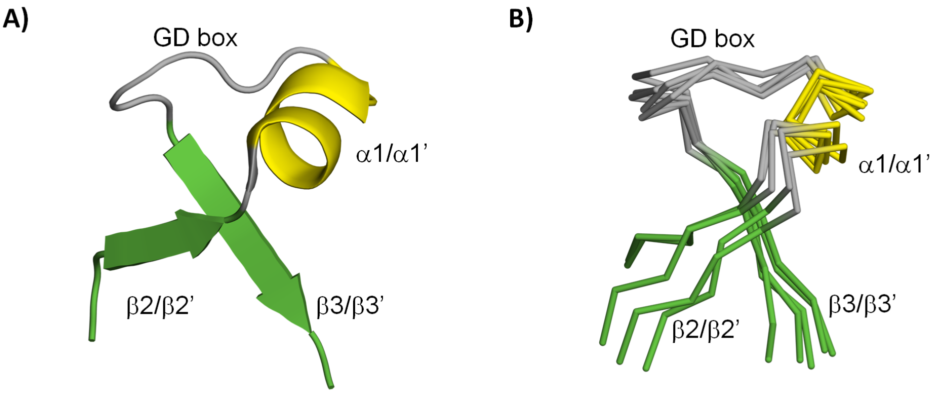


Figure 6.1: Gallery of cradle-loop barrels. In all structures, β-strands are colored in green, cradle loops in black, helices in yellow, β-strands inserted into the barrel in blue, and elements not part of the barrel in gray. The $\beta 2$ and $\beta 2'$ strands in the RIFT and double-psi barrels are shown in red and pink, respectively. In homodimeric structures, monomers are distinguished by light and dark colors. The structures shown are: A) double-psi barrel: VatN-N (1CZ4), B) RIFT barrel: PhS018 (2GLW), C) swapped-hairpin barrel: AbrB (homodimeric barrel, 1YFB), D) swapped-hairpin barrel: MraZ (monomeric barrel, 1N0E), E) C-terminal domain of bacterial fluorinating enzyme (1RQP), and F) C-terminal domain of AstA (1YLE).



C)

1CZ4(A:22-46)	double-psi	**S**R**V**R**L**DESS**R**RL**L**D**A**EI**GD**V**VEI**EK
1CZ4(A:65-89)	double-psi	**G**I**V**R**I**DSVM**R**NN**CGA**SI**GD**K**V**K**V**RK
2GLW(A:15-39)	RIFT	**G**R**I**I**I**PAGT**R**KF**YGI**EQ**GD**F**VEI**KI
2GLW(A:60-84)	RIFT	**G**S**V**I**I**PKAL**R**DV**IGI**KP**GE**V**IEV**LL
1YFB(A:16-40)	swapped-hairpin	**G**R**V**V**I**PIEL**R**RT**LGI**AEK**D**A**LEI**YV

Figure 6.2: The conserved β-α-β motif. In all structures, β-strands are colored in green, loops in gray, and α-helices in yellow. A) The first β-α-β motif from VatN-N (1CZ4) is shown in cartoon representation. B) Superposition of the β-α-β motifs from VatN-N (1CZ4, both repeats), PhS018 (2GLW, both repeats), and AbrB-N (1YFB) is shown in backbone representation. C) Sequence alignment corresponding to the structural alignment shown in panel B. Conserved residue columns are indicated in blue.

particular interest were dimeric bacterial transcription factors, typified by *Bacillus subtilis* AbrB, whose N-domains carried a single GD box and were elaborated by an additional C-terminal β-strand. They thus appeared to represent a homodimeric precursor to the double-psi barrel, with a permuted fold that would resolve the topological complexities of the latter. This led us to propose this topology as the ancestral form [Coles et al., 1999]. Our hypothesis was contradicted by the published structure of AbrB-N (1EKT), which, instead of a barrel, showed a side-by-side dimer of two three-stranded β-meanders with two equatorial helices.

In order to understand this contradiction, we investigated a group of sequences that were intermediate between VatN-N and AbrB. Proteins of this group, typified by PhS018 from *Pyrococcus horikoshii*, either have internal sequence symmetry and carry two copies of the GD box or are homodimers with one copy per subunit. PhS018 turned out to have yet a third fold, forming a singly interdigitated, six-stranded barrel (Fig. 6.1B) [Coles et al.,

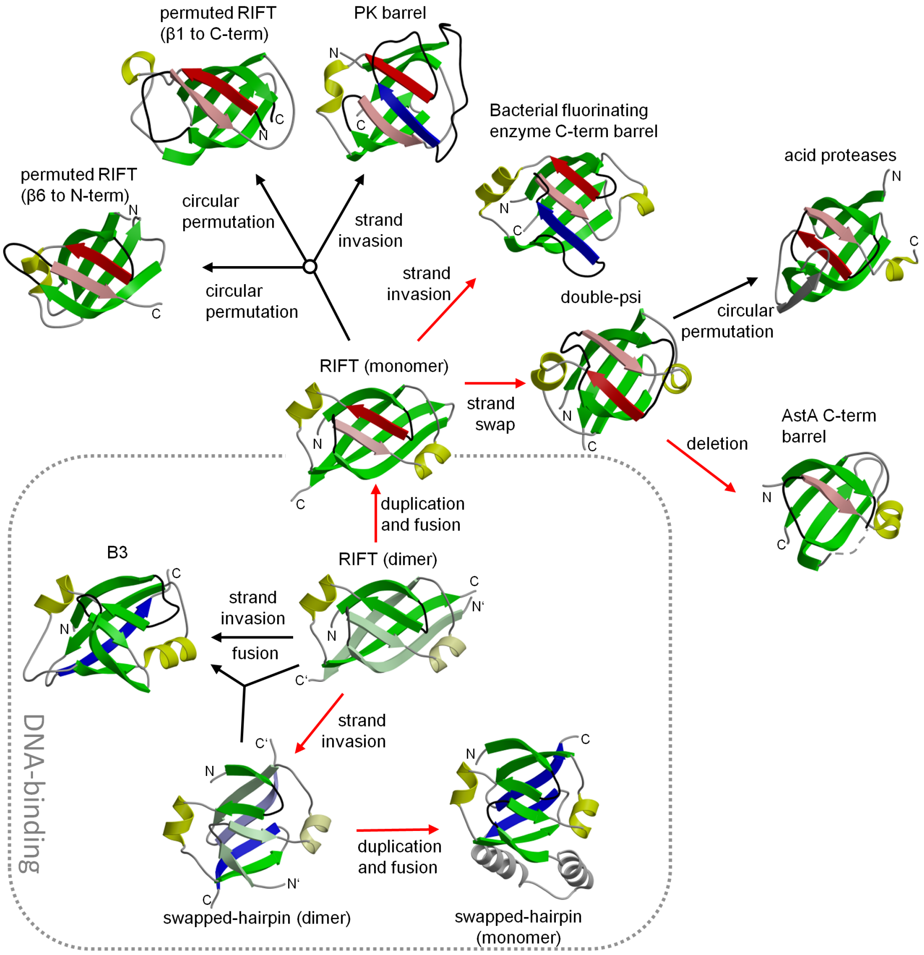


Figure 6.3: Topological map of the cradle-loop barrel metafold (modified from [Alva et al., 2008]). In all structures, β-strands are colored in green, cradle loops in black, helices in yellow, β-strands inserted into the barrel in blue, and elements not part of the barrel in gray. The $\beta2$ and $\beta2'$ strands in the RIFT and double-psi barrels are shown in red and pink, respectively. In homodimeric structures, monomers are distinguished by light and dark colors. Homologous connections, as substantiated by HMM-HMM comparisons [Soeding, 2005], are indicated by red arrows. The structures shown are: I) swapped-hairpin barrels: AbrB (homodimeric barrel, 1YFB) and MraZ (monomeric barrel, 1N0E); II) RIFT barrels: PhS018 (monomeric barrel, 2GLW) and the homology model of MTpME2200 Orf5 based on PhS018 (homodimeric barrel), EF-Tu ($\beta1$ to C-terminus, 1D2E), V8 protease ($\beta6$ to N-terminus, 1QY6), the PK barrel (1A49), B3 barrel (1WID), and the C-terminal domain of bacterial fluorinating enzyme (1RQP); and III) double-psi barrels: VatN-N (1CZ4), the C-terminal domain of AstA (1YLE), and HIV-1 protease (1NH0). Details of the structures are given in Tables 6.1 and 6.2.

2006]. We named this topology the RIFT barrel for its widespread occurrence in ancient proteins, such as the ribosomal protein L3, the N-domain of the F1 ATPase, and translation factors of the EF-Tu family. This topology was clearly related to that of double-psi barrels by a strand swap of the symmetry-related $\beta2/\beta2'$ strands, but was not visibly related to the published AbrB fold, 1EKT. In particular, the conserved GD boxes resembled closely those of VatN-N, but had an entirely different conformation from 1EKT.

Given this discrepancy, we decided to redetermine the structure of AbrB-N, which turned out to neither resemble the published structure (which was subsequently retracted), nor a permuted form of the double-psi barrel. Rather, additional C-terminal strands were inserted into the RIFT barrel to form an eight-stranded architecture with two pairs of interdigitated β-hairpins (Fig. 6.1C), leading us to name this fold the swapped-hairpin barrel [Coles et al., 2005]. Significantly, the GD-box region now resembled the equivalent regions closely in double-psi and RIFT barrels.

Our results on VatN-N, PhS018, and AbrB-N confronted us with a problem of nomenclature, as the three barrels were clearly homologous, but equally clearly had different folds. In search of a term that would describe their relationship, we chose to define them as a metafold and denote them as cradle-loop barrels for the distinctive profile conferred by their characteristic $\beta1$-$\beta2$ loops.

Thus, in our evolutionary scenario, an ancestral, homodimeric RIFT barrel gave rise to swapped-hairpin barrels by strand invasion and to double-psi barrels by fusion and strand swapping (Fig. 6.3). We also found sequence similarity indicative of homology between RIFT barrels and yet a fourth fold, the C-terminal domain of bacterial fluorinating enzyme: this is related to the RIFT topology by a strand invasion from the second cradle loop into the space between $\beta1'$ and $\beta2'$, yielding a seven-stranded barrel (Fig. 6.1E). Jointly, these proteins map out a network of homologous but topologically distinct folds (Fig. 6.3).

6.3.2 The peripheral group

We placed the RIFT barrel at the center of the cradle-loop network (Fig. 6.3), because of its simpler topology and its occurrence in a wide range of ancient proteins. We included all proteins with a RIFT barrel fold as candidate groups, even though we initially had no evidence for their homologous relationship to the proteins we had used to define the metafold. We did this in order to map out the possible relationships that would be explored next most usefully. Recently, we substantiated a homologous relationship between the core group and one candidate group, the riboflavin kinases (SCOP b.43.5, see Table 6.1), by identifying a family of archaeal proteins that bridge the evolutionary space between the two. These proteins are similar in sequence

Table 6.1: RIFT barrel proteins (modified from [Alva et al., 2008])

SCOP	Superfamily[a b]	PDB code[c] (chain, residues)
Homodimeric barrel		
-	Af2212 from *A. fulgidus*	2NWT (A, B)
Monomeric barrel		
-	PhS018 from *P. horikoshii**	2GLW
b.43.2	FucI/AraA C-term domain-like	1FUI (A:356-591)
b.43.3	Translation proteins+	1EFC (A:205-296)
b.43.4	Riboflavin synthase domain-like*	1I8D (A:6-92)
b.43.5	Riboflavin kinase-like*	1N08
b.49.1	N-domain of F1 ATP synthase α and β subunits	1W0J (A:28-91)
b.49.2	Alanine racemase C-ter domain-like*	1BD0 (A:243-336)
b.49.3	Aminopeptidase/glucanase lid domain	1VHE (A:73-163)
b.129.2	PG0164-like*	2D9R (A:20-104)
e.56.1	YaeB-like	1XQB (A:1-63, 99-136)
f.46.1	HlyD-like secretion proteins*	1VF7 (A:26-37, 173-226)
β1 circularly permuted to C-terminus		
b.44.1	EF-Tu/eEF-1/eIF2- C-term domain	1D2E (A:349-451)
b.44.2	Aminomethyltransferase β-barrel+	1WOS (A:279-361)
β6 circularly permuted to N-terminus		
b.45.1	FMN-binding split barrel	1FLM (A)
b.45.2	PilZ domain-like+	1YLN (A:138-248)
b.47.1	Trypsin-like serine proteases	1QY6 (A)
b.106.1	Phage tail proteins+	1K0H (A)
b.140.1	Replicase NSP9	1QZ8 (A)
e.53.1	QueA-like	1VKY (A:64-142)
β-strand inserted from cradle-loop 1		
b.58.1	PK β-barrel domain-like+	1A49 (A:116-217)
β-strand inserted from cradle-loop 2		
b.141.1	Bacterial fluorinating enzyme, C-domain*	1RQP (A:193-298)
Additional C-terminal β-strand		
b.142.1	DNA-binding pseudobarrel domain	1WID (A)

[a] Homology between superfamilies was evaluated with HHsearch [Soeding, 2005].
[b] Superfamilies showing sequence similarity indicative of homology fall into two separate networks, marked with (*) and (+), as described in the text.
[c] Representative structures are shown in Fig. 6.1 and Fig. 6.3.

Table 6.2: Swapped-hairpin barrel and Double-psi barrel proteins

SCOP	**Superfamily**[a]	**PDB code (chain, residues)**
Swapped-hairpin barrel		
Homodimeric barrel		
b.129.1	AbrB/MazE/MraZ-like*	1MVF (D, E)
Monomeric barrel		
b.129.1	AbrB/MazE/MraZ-like*	1N0E (A)
Double-psi barrel		
b.52.1	Barwin-like endoglucanases	2ENG
b.52.2	ADC-like*	1CZ4 (A: 1-91)
e.29.1	β and β' subunits of DNA dept. RNA-pol	1SMY (C: 668-698, 832-1004)
$\beta 6$ circularly permuted to N-terminus		
b.50.1	Acid proteases	1NH0 (A)
$\beta 2$ and $\alpha 1$ deleted		
d.108.1	Acyl-CoA N-acyltransferases*	1YLE (A:273-340)

[b] Superfamilies marked with a (*) are part of the (*) network in Table 6.1.

to both core and candidate group, and have CTP-dependent riboflavin kinase activity [Ammelburg et al., 2007].

In order to obtain a more complete view of the potential homologous protein space, we think that candidate groups would also usefully include proteins that have at most one topological change relative to the core group, provided that this change is known to occur between homologous proteins. Our current list of cradle-loop barrels includes a series of candidate topologies, related in this way to the three core topologies: RIFT, swapped-hairpin, and double-psi (Fig. 6.3; Tables 6.1 and 6.2).

For the RIFT barrel, two variants are obtained by circular permutation and three by the insertion of an additional strand into the barrel. The elongation factor/aminomethyl-transferase common domain (SCOP b.44) is formed by the circular permutation of $\beta 6$ to the N-terminus and the FMN-binding split barrel (b.45), trypsin-like serine proteases (b.47), replicase NSP9 (b.104), phage tail proteins (b.106), and the QueA-like barrel (e.53) by the circular permutation of $\beta 1$ to the C-terminus. In the PK β-barrel domain (b.58), the barrel is elaborated by the insertion of a β-strand originating from the first cradle loop, and in the aforementioned C-terminal domain of bacterial fluorinating enzyme (b.141) by the insertion of a β-strand from the second cradle loop. In the DNA-binding pseudobarrel (b.142), the barrel is extended by the insertion of an additional C-terminal strand. This barrel might have been formed either by strand insertion, or by the fusion of two different half barrels, the RIFT monomer and the swapped-hairpin monomer. In the RIFT fold space, we are observing the emergence of a second homologous network, based on translation proteins (b.43.3) and including members of three variant RIFT folds (b.44.2, b.45.2, b.106.1, b.58.1), as listed in Table 6.1. We have as yet no evidence for the homology of this network to our core cradle-loop network.

For the double-psi barrel, one variant arises by circular permutation of $\beta 6$ to the N-terminus, as seen in acid proteases (b.50), and one by the deletion of $\beta 2$ and $\alpha 2$, as seen in acyl-CoA N-acyltransferases (d.108). Although the acyltransferases have clear sequence similarity to double-psi barrels, aspartic proteases do not. Nevertheless, their inherent chaperone-like activity may point to a distant evolutionary connection with the double-psi barrel domains of AAA-ATPases [Hulko et al., 2007].

We do not currently have fold variants for the swapped-hairpin barrels, but we note that some monomeric proteins of this group have lost strand $\beta 1'$ [Coles et al., 2005], pointing to further topological variants in the network.

6.4 Moving towards a Linnaean-like classification system for proteins

The combination of sequence and fold similarity in generating current protein classifications, that is of homologous and analogous criteria, introduces contradictions between and within systems. We see the metafold as a useful concept for addressing contradictions arising from homologous fold change and from the detection of homologous fragments in domains of different folds. Although primarily sequence-driven and thus based on homology, the metafold concept still uses fold similarity in order to identify candidate folds. Ultimately, however, contradictions can only be addressed comprehensively by eliminating all analogous classification criteria. At that point, we would obtain a classification of proteins by natural descent, conceptually related to the Linnaean system for organisms, albeit with multiple roots, as proteins are not monophyletic. We think that such a system is highly desirable, as homology offers a rich source of structural, functional, and mechanistic information, while analogy is comparatively uninformative and often misleading. Indeed the central role of model systems in modern biology can only be understood in terms of extrapolation by homology, since few researchers would be interested in *Danio*, *Drosophila*, or *Caenorhabditis* for their own sake.

If a classification by natural descent is so desirable, why has analogy played such a great role in all classification efforts so far? We would argue that this was by default, as homologous criteria were not available: sequence search methods were not sufficiently developed to reveal remote homology, sequence databases were too sparse to allow for efficient connections in sequence space, and there were too few structures to validate distant relationships. Also, proteins are not monophyletic, so that - unlike in the Linnaean system - analogy was the default assumption for observed similarities. This situation is changing. With the emergence of profile search methods and, more recently, with methods based on the comparison of profile Hidden Markov Models [Soeding, 2005], bioinformatic tools have reached considerable sensitivity. Sequence databases have been growing fairly steadily by about one order of magnitude every five years, and currently contain about 1.4×10^7 proteins [GenBank, 2012; RefSeq, 2012] . Given a global proteome of about one trillion proteins ($\approx10^8$ species with $\approx10^4$ protein-coding genes each), at current rates we might know the sequence of most proteins on Earth in about a quarter century. Of possibly greater immediate impact, there are now some 3,100 complete genomes from all over the tree of life, also growing by about one order of magnitude every five years [GOLD, 2012]. In parallel, the number of structures known to atomic resolution, currently at about 8×10^4, has been growing steadily, if more slowly, by about one order of magnitude every seven years [PDB, 2012](the number of non-redundant structures with at most 30% pairwise sequence identity is about one-fifth this size). Given that we currently recognize $\approx10^4$ protein families [Hunter et al., 2009; Punta et al., 2012] and this number is unlikely to rise by more than at most one order of magnitude (if indeed it will rise at all), it seems likely that most protein families will have at least one member of known structure within the next 10 to 20 years [Marsden et al., 2007]. By targeting proteins from a wide range of families that have remained unexplored, without regard to the availability of functional information, structural genomics initiatives are playing a key role in this effort [Burley and Bonanno, 2002; Chandonia and Brenner, 2006]. The wide availability of sequence and structure data for most families is essential for substantiating remote homology, since sequence similarity as a function of structure similarity is a powerful discriminator between homology and analogy [Remmert et al., 2009]. We therefore wish to argue that it has become possible to start removing analogous criteria across protein classifications and move toward a system based on natural descent.

6.5 Materials and methods

All sequence similarity searches were carried out in the MPI bioinformatics toolkit [Biegert et al., 2006] using HHpred [Soeding, 2005] with default settings and a probability cutoff of 40%. HHpred searches were performed against SCOP70, which is a version of SCOP (version 1.73) [Andreeva et al., 2004] filtered for a maximum of 70% pairwise sequence identity. We seeded our searches with the following representative structures: double-psi barrel (VatN-N, 1CZ4), RIFT barrel (PhS018, 2GLW), and swapped-hairpin barrel (AbrB, 1YFB). The resulting matches were analyzed interactively and folds related by homologous fold change events were pooled together. The validity of the obtained relationships were confirmed by comparing them to previously published reports [Coles et al., 1999, 2005, 2006; Ammelburg et al., 2007; Hulko et al., 2007]. To detect homologs that may not have been classified into SCOP, additional searches were carried out against the PDB70 database using HHpred. To search for folds related only in structure, we made pairwise comparisons of the aforementioned representative structures with every structure in the SCOP70 dataset using the structural alignment method TMalign [Zhang and Skolnick, 2005]. We pooled together all matches with a RMSD of less than 5Åand analyzed them interactively.

Chapter 7

Summary and conclusions

In this dissertation, we took advantage of the recent growth of protein sequence and structure databases, and the progress made in profile-based sequence comparison methods to study the polyphyly, origin, and classification of proteins.

Despite their enormous sequence diversity, modern proteins are built of a limited set of recurrent domains belonging to only about 10000 homologous families, which in turn form about 3000 broader evolutionary superfamilies [Hunter et al., 2009]. This limited diversity of proteins is even more restricted at the structural level, as many superfamilies have the same fold, even in cases when they have no obvious evolutionary relationship. In fact, the 3000 domain superfamilies seen in modern proteins can be assigned to one of about 1000 folds based on the similarity of their tertiary structures [Andreeva et al., 2008]. In the absence of detectable sequence similarity, the structural similarities between superfamilies of one fold were long thought to have originated independently, by convergent evolution. However, the recent growth of molecular databases and progress in sequence comparison methods have brought forward numerous cases of distant evolutionary relationships between superfamilies of one fold, suggesting that folds may not be as polyphyletic as hitherto assumed. Nevertheless, the pervasiveness of such relationships in the fold space is still unclear. To investigate this, we clustered representative domains of known fold types by their sequence similarity, as evaluated by the-state-of-the-art remote homology detection method HHsearch [Soeding, 2005], and produced a two-dimensional map of the fold space (*galaxy of folds*), with a high-level view of the evolutionary relationships in it (Chapter 3). As expected, families belonging to the same superfamily form tight clusters. But frequently, superfamilies of the same fold are also connected to each other, making clear that, far from being anecdotal, these relationships are widespread and that many folds are monophyletic rather than have independent origin. For instance, our map shows that, of the 25 folds with the most number of superfamilies in the

SCOP database [Andreeva et al., 2008], 16 contain superfamilies that either show homologous connections or a mixture homologous and analogous connections. Although our map reveals that folds may not be as polyphyletic as considered by SCOP, we do find instances of analogous folds, the most interesting example of which is the ferredoxin-like fold, which has the largest number of superfamilies in SCOP. In sum, our galaxy of folds provides new insights into the organization of the protein universe, by capturing most known and many previously unknown evolutionary relationships between protein superfamilies.

Despite the wide acceptance that the number of domains or folds in nature is limited, the origin of this set of folds itself is poorly understood. One compelling theory for the origin of folded domains proposes that they arose by fusion and recombination from an ancestral set of peptides, which emerged as cofactors in the context of the RNA world [Lupas et al., 2001; Soeding and Lupas, 2003]. According to this model, the assemblage of these peptides led to protein folding as an emergent property, based on the ability of the peptides to exclude water between each other, rather than with the RNA scaffold. We reasoned that if this hypothesis is correct, we should still be able to see traces of these peptides in modern proteins. To this end, we compared domains of all known fold types using a sequence- and structure-based approach, and detected 50 homologous fragments that occur in domains with different folds (Chapter 3). A third of the fragments we identified had been noticed individually before, confirming the validity of our approach. These 50 potential ancestral fragments are subdomain-sized, comprising two or three secondary structure elements, and are widespread across fold types, especially across populous folds. This provides strong evidence for our proposition that domains arose from simpler fragments. Based on their involvement in the most ancient functions, e.g. nucleic acid-binding and metal-binding, on their occurrence in the most ancient folds, e.g. the P-loop-containing nucleoside triphosphate hydrolase fold and the DNA/RNA-binding three-helical bundle fold, and on the resemblance of their shapes to that of ribosomal proteins, we propose that the 50 fragments we identified may represent the observable remnants of the RNA-peptide world from which the first folded domains arose. Intriguingly, as opposed to our initial expectation, we did not find a single case of a domain containing two different fragments from our set of ancestral peptides, indicating that assembly from non-identical fragments may not have been an important mechanism in the emergence of domains. However, we do find many examples of domains that are built from the same fragment either by accretion, that is, by addition of multiple structural elements at the ends, or by amplification, that is, by repeating the same unit multiple times. This reveals that accretion and repetition were the primary factors in the emergence of domains. In sum, our findings indicate that the first, ancestral proteins were subdomain-sized peptides with minimal structure and basic function. These peptides then evolved to more stable domains by either amplification or by accretion. The multidomain proteins, that we recognize today, arose subsequently from domains by recombination, amplification, and divergence.

In Chapter 4, we retraced the evolutionary events that may have led to the emergence of histones, and detected homologous connections to two other domains, the N-terminal substrate recognition domain of Clp/Hsp100 proteins (N-domain) and the helical part of the extended AAA+ ATPase domain (C-domain). The similarity between these three domains is hinged on the presence of a shared ancestral fragment corresponding to a helix-strand-helix motif. We propose that this motif first gave rise divergently to the N-domain of Clp/Hsp100 proteins and to the C-domain of AAA+ proteins. The histone fold arose subsequently from the C-domain through a domain-swapping event. Although it had previously been proposed that domain swapping might contribute to protein evolution [Kinch and Grishin, 2002], no naturally-occurring examples were known. This is, to the best of our knowledge, the first example of a genetically fixed three-dimensional domain swap that resulted in the creation of a protein family with a novel fold. A recent study has further strengthened our hypothesis by providing experimental evidence for the origin of the histone fold through a domain-swapping event [Hadjithomas and Moudrianakis, 2011].

Modern proteins are very diverse and ordering them into a taxonomy based on natural descent, similar to the taxonomy of organisms, would be of great benefit, as homology provides an informative resource for the comparative studies for proteins. Such a classification system is, however, presently unavailable. Current schemes classify proteins by combining homologous criteria at lower classification levels with analogous criteria at upper levels. By doing so, these systems make the implicit assumption that homologous domains always have the same fold. However, as has become evident in recent years, this is not always the case. Over the course of evolution, homologous domains may evolve different structures, owing to events such as circular permutation. Also, as discussed in this thesis, domains of different folds may be evolutionarily related because of emergence from a shared ancestral peptide. In Chapter 6, we argued that, with the recent growth in molecular databases and the improved sensitivity of profile-based remote homology detection methods, it should now become possible to detect and unify protein superfamilies linked by such distant evolutionary relationships. To this end, we introduced a new classification level, the metafold, which captures groups of folds that are related by homologous descent. We used cradle-loop barrels [Coles et al., 2005, 2006], a group of topologically different but homologous folds, to put forward our ideas on the metafold. We expect that the metafold will prove to be an important step on the way to a classification system for proteins based on natural descent.

s

Appendix A

Ancestral peptides

PDB accession	SCOP ID	Alignment
		hhhhhhhhcc-chhhhhhhhh
1HLV (A:28-47)	a.4.1.7	KGE**I**ARR**F**N**I**-PPST**L**ST**I**LK
1MGT (A:112-131)	a.4.2.1	YGD**L**AKA**L**NT-SPRA**V**GG**A**MK
1J5Y (A:27-46)	a.4.5.1	GAQ**L**AEE**L**S**V**-SRQV**I**VQDIA
1FSE (A:29-48)	a.4.6.2	TKE**I**ASE**L**F**I**-SEKT**V**RNHIS
1JHG (A:68-87)	a.4.12.1	QRE**L**KNE**L**G**A**-GIAT**I**TR**G**SN
1KU3 (A:398-417)	a.4.13.2	LEE**V**GAY**F**G**V**-TRER**I**RQ**I**EN
1VZ0 (A:137-156)	a.4.14.1	QEE**V**ARR**V**GK-ARST**V**AN**A**LR
1V92 (A:9-28)	a.5.2.3	LRE**F**VAVTG**A**-EEDR**A**RF**F**LE
1R8E (A:8-27)	a.6.1.3	IGE**V**SKL**A**N**V**-SIKA**L**RY**Y**DK
2R1J (A:21-40)	a.35.1.2	QAA**L**GKM**V**G**V**-SNVA**I**SQ**W**ER
2CPG (A:17-36)	a.43.1.3	LEK**M**ARE**M**G**L**-SKSA**M**IS**V**AL
1AIS (A:1268-1287)	a.74.1.2	QRE**V**AEV**A**R**V**-TEVT**V**RNRYK
1A9X (A:499-518)	a.92.1.1	DAR**L**AKL**A**G**V**-REAE**I**RK**L**RD
1Z67 (A:75-94)	a.259.1.1	VSD**L**GQK**L**G**V**-DTST**A**SS**L**LA
2CSB (A:200-219)	a.267.1.1	HDE**I**ARR**L**G**L**-SVSE**V**EGEKD
1BOB (A:293-312)	d.108.1.1	LESSRKS**L**K**L**-EERQ**F**NR**L**VE
1F44 (A:306-326)	d.163.1.1	IPE**I**MQA**G**G**W**TNVNI**V**MN**F**IR
2V4J (C:66-85)	d.203.1.1	VRI**L**SKNTG**F**-KLKE**V**YE**L**FP
1NR3 (A:8-27)	d.236.1.1	QKK**I**ARE**L**KT-TRQN**V**SA**I**ER
1I3J (A:215-234)	d.285.1.1	AAD**A**ARH**F**K**I**-SSGL**V**TYRVK
1MW9 (A:290-309)	e.10.1.1	QQA**A**STR**L**G**F**-GVKKTMM**M**AQ
1LDJ (A:622-641)	e.40.1.1	VQQ**L**TDSTQ**I**-KMDI**L**AQ**V**LQ
2HQ2 (A:19-38)	e.62.1.1	ARD**I**AGL**M**N**I**-REAE**L**AF**A**RV
2AVU (E:26-45)	e.64.1.1	LQM**L**ESETQ**L**-SRGR**L**IK**L**YK

Table A.1: DNA-binding helix-turn-helix (HTH) motif (Fragment 1). Found in 18 folds comprising 24 superfamilies. This motif has been previously described [Brennan and Matthews, 1989; Brennan, 1993; Aravind et al., 2005]. We detected this fragment at HHsearch probabilities ranging between 50%-90%. The HTH motif binds DNA in a sequence specific manner. It contains two short helices connected by a short turn. While the second helix is involved in interactions with DNA, the first helix stabilizes the structure. Although the HTH motif of 1NR3 (SCOP d.236.1.1), which was solved using NMR spectroscopy, makes matches in sequence to other HTHs, its structure is non-canonical. We think that this could be an experimental artifact. We cannot confirm this presently because there are no alternative structures of this protein or its orthologs in the PDB. In the alignment, conserved residue columns are shown in bold face and the secondary structure composition is indicated above the alignment.

PDB accession	SCOP ID	Alignment
		hhhhh------cccccchhhhhhhhhhh
1IXR (A:71-92)*	a.60.2.1	FEL**L**L------S**V**S**GVG**PKV**A**LA**L**LSAL
1IXR (A:106-127)*	a.60.2.1	ARL**L**T------S**A**S**GVG**RRL**A**ER**I**ALEL
1z3e (B:282-303)	a.60.3.1	EED**M**M------K**V**RN**LG**RKS**L**EE**V**KAKL
2I1Q (A:35-56)	a.60.4.1	VGE**L**T------D**I**E**GI**SEKA**A**AK**M**IMGA
1CI4 (A:17-38)	a.60.5.1	EKP**V**G------S**L**A**GIG**EVL**G**KK**L**EERG
2FMP (A:56-77)	a.60.6.1	GAE**A**K------K**L**P**GVG**TKI**A**EK**I**DEFL
1RXW (A:228-255)	a.60.7.1	IAI**L**VgtdyneG**V**K**GVG**VKK**A**LN**Y**IKTY
1WUD (A:574-595)	a.60.8.1	ASE**M**L------S**V**N**GVG**MRK**L**ER**F**GKPF
2FMP (A:97-118)	a.60.12.1	INF**L**T------R**V**S**GIG**PSA**A**RK**F**VDEG
1ORN (A:109-130)	a.96.1.1	RDE**L**M------K**L**P**GVG**RKT**A**NV**V**VSVA
2GY9 (M:15-36)	a.156.1.1	VIA**L**T------S**I**Y**GVG**KTRSKA**I**LAAA
2P6R (A:631-652)	a.289.1.2	LLE**L**V------R**I**RH**IG**RVR**A**RK**L**YNAG
1GM5 (A:114-135)	b.40.4.9	STD**I**Q------Y**A**K**GVG**PNRKKK**L**KKLG
1JX4 (A:177-198)	e.8.1.7	ELD**I**A------D**V**P**GIG**NIT**A**EK**L**KKLG
1VDD (A:11-32)	e.49.1.1	IRE**L**S------R**L**P**GIG**PKS**A**QR**L**AFHL
2I5H (A:129-150)	e.71.1.1	MHQ**L**E------L**L**P**GVG**KKM**M**WA**I**IEER

Table A.2: Helix-hairpin-helix (HhH) motif (Fragment 2). Found in 8 folds comprising 15 superfamilies. This motif has been previously described [Doherty et al., 1996]. We detected this fragment at HHsearch probabilities ranging between 60%-95%. The HhH motif is found in non-sequence specific DNA-binding proteins. It contains two helices connected by a short loop and contacts to DNA are made by amino acids located in this loop. RuvA domain 2-like superfamily (a.60.2.1) contains two homologous copies of this motif; the copies are indicated by a *.

PDB accession	SCOP ID	Alignment
		hhhhhhhhhhhhhcchhhhhhhhhhhhhh
2AHO (B:101-130)	a.60.14.1	QR**L**D**KIL**EL**VSQK**LKLSEK**DAW**E**QVA**WKLE
1U55 (A:46-75)	d.278.1.1	DE**V**R**RIF**AK**VSEK**TGKNVN**EIW**R**EVG**RQNI

Table A.3: α-hairpin from eIF2alpha (Fragment 3). Found in 2 folds comprising 2 superfamilies. We detected this fragment at HHsearch probabilities ranging between 60%-70%.

PDB accession	SCOP ID	Alignment
		hhhhhhhhhhhhhhhccchhhhhhhhhhhhh
2HEP (A:6-36)	a.2.21.1	KI**AR**IN**ELAAK**A**K**AG**VIT**E**EEKAE**QQK**L**R**QE**
1T3W (A:547-577)	a.236.1.1	LE**LR**QE**ELIAR**E**R**TH**GLS**N**EERLE**LWT**L**N**QE**

Table A.4: α-hairpin from DnaG, C-terminal domain (Fragment 4). Found in 2 folds comprising 2 superfamilies. We detected this fragment at HHsearch probabilities ranging between 90%-95%.

PDB accession	SCOP ID	Alignment
		hhhhhhhhcccchhhhhhhhhhhhh
1EI1 (A:354-377)	d.14.1.3	NEL**LAEYL**L**EN**PTD**A**KI**V**VGK**I**ID
1S16 (A:1351-1374)	d.14.1.3	KDA**FILWL**N**QN**VQA**A**EL**L**AEM**A**IS
1U94 (A:302-325)	d.48.1.1	KAN**ATAWL**K**DN**PET**A**KE**I**EKK**V**RE

Table A.5: α-α motif from RecA, C-terminal domain (Fragment 5). Found in 2 folds comprising 2 superfamilies. We detected this fragment at HHsearch probabilities ranging between 50%-55%.

PDB accession	SCOP ID	Alignment
		hhhhhhhhhhhhcccchhhhhhhhhh
1XB2 (B:59-83)	a.5.2.2	SKE**LL**MK**L**RRK**TGYS**FI**N**CKK**ALE**T
1AIP (C:3-27)	a.5.2.2	QME**LI**KK**L**REA**TGAG**MM**D**VKR**ALE**D
1EFU (B:4-28)	a.5.2.2	TAS**LV**KE**L**RER**TGAG**MM**D**CKK**AL**TE
1B8Z (A:2-26)	a.55.1.1	KKE**LI**DR**V**AKKA**GA**KKK**D**VKL**ILD**T
1HUU (A:3-27)	a.55.1.1	KTE**LI**NA**V**AET**SGLS**KK**D**ATK**AVD**A
1CTF (A:65-89)	d.45.1.1	KVA**VI**KA**V**RGA**TGLG**LK**E**AKD**LVE**S
1DD3 (A:70-94)	d.45.1.1	KIQ**VI**KV**V**REI**TGLG**LK**E**AKD**LVE**K

Table A.6: EF-Tu binding α-hairpin (Fragment 6). Found in 3 folds comprising 3 superfamilies. This motif has been previously described [Wieden et al., 2001]. We detected this fragment at HHsearch probabilities ranging between 60%-85%. In EF-Ts (a.5.2.2) and Ribosomal protein L7/12 (d.45.1.1) this motif has been implicated to be involved in interactions with EF-Tu.

PDB accession	SCOP ID	Alignment
		hhhhhhhhhhhhhhccccchhhhhhhhhhh
1K78 (A:91-120)	a.4.1.5	**ATP**K**VV**EK**IA**E**Y**KRQN**P**TMFAW**EIRDRL**LA
6PAX (A:76-105)	a.4.1.5	**ATP**E**VV**SK**IA**Q**Y**KQEC**P**SIFAW**EIRDRL**LS
1GCI (A:223-252)	c.41.1.1	**ATP**H**VA**GA**AA**L**V**KQKN**P**SWSNV**QIRN**H**L**KN
1R0R (E:223-252)	c.41.1.1	**ASP**H**VA**GA**AA**L**I**LSKH**P**NLSAS**QVRNRL**SS

Table A.7: Helix-turn-helix motif from subtilisin (Fragment 7). Found in 2 folds comprising 2 superfamilies. We detected this fragment at HHsearch probabilities ranging between 60%-70%. In subtilisins (c.41.1.1), this motif contains the catalytic serine/threonine. In the paired domain (a.4.1.1), it mounts the DNA-binding helix.

PDB accession	SCOP ID	Alignment
		hhhhhhhchhhhhhhhhhhh
1VQO (V:4-23)	a.2.2.1	HVQ**EIR**D**M**TPA**E**REAE**L**DDL
1Y02 (A:72-91)	a.140.2.1	QRE**EL**MK**M**KVK**D**LRDY**L**SLH
1A62 (A:2-21)	a.140.3.1	NLT**ELK**NTPVS**E**LITL**G**ENM
2HJQ (A:58-77)	a.140.3.1	TES**ELK**G**M**NKA**E**HESI**I**SNL

Table A.8: α-α-motif from ribosomal protein L29 (Fragment 8). Found in 2 folds comprising 3 superfamilies. We detected this fragment at HHsearch probabilities ranging between 60%-70%.

PDB accession	SCOP ID	Alignment
		hhhhhhhhhhhhhhhccccceechhhhhhhh
1K6K (A:5-34)*	a.174.1.1	E**L**ELS**L**NM**A**FAR**A**REHRHEF**M**TVEH**L**LL**A**L
1K6K (A:83-112)*	a.174.1.1	S**F**QR**VL**QR**A**VFH**V**QSSG**R**NE**V**TGAN**V**LV**A**I
1TZY (A:58-87)	a.22.1.1	L**T**AE**IL**EL**A**GNA**A**RDNK**K**TR**I**IPRH**L**QL**A**I
1TAF (A:52-82)	a.22.1.3	Y**V**TS**IL**DD**A**KVY**A**NHAR**K**KT**I**DLDD**V**RL**A**T
1FNN (A:241-270)	c.37.1.20	L**A**ID**IL**YR**S**AYA**A**QQNG**R**KH**I**APED**V**RKSS
1LV7 (A:363-392)	c.37.1.20	D**L**AN**LV**NE**A**ALF**A**ARGN**K**RV**V**SMVE**F**EK**A**K

Table A.9: Helix-strand-helix (HSH) motif (Fragment 9). Found in 3 folds comprising 3 superfamilies. We have previously described this motif [Alva et al., 2007]. We detected this fragment at HHsearch probabilities ranging between 50%-90%. The HSH motif is implicated to be involved in binding DNA and in protein-protein interactions. The N-terminal substrate recognition domain of Clp/Hsp100 proteins contains two homologous copies of this motif (indicated by a * in the alignment).

PDB accession	SCOP ID	Alignment
		hhhhhhhhhhhhhhhhhhhccchhhhhhhhhhhhhhhhh
1SR2 (A:813-850)	a.24.10.4	LFVDT-**VP**DD**V**KR**L**YTE**A**ATS**D**FAA**LA**QTAHR**L**KGV**F**AM
2A0B (A:686-723)	a.24.10.1	VFEKM-**MP**GY**V**SV**L**ESN**L**TAQ**D**KKG**IV**EEGHK**I**KGA**A**GS
1FI4 (A:224-262)	d.58.26.2	ERIEHV**VP**KR**F**EV**M**RKA**I**VEK**D**FAT**FA**KETMM**D**SNS**F**HA

Table A.10: α-hairpin from GHMP kinase, C-terminal domain (Fragment 10). Found in 2 folds comprising 2 superfamilies. We detected this fragment at HHsearch probabilities ranging between 55%-60%. The HTP domain (a.24.10) contains two structurally similar copies of this motif, but they do not exhibit any detectable sequence similarity.

PDB accession	SCOP ID	Alignment
		hhhhhhhcchhhhhhhhhhhc-eeeeeec
1TUA (A:13-40)	d.51.1.1	RL**G**A**VIG**PR**G**EVKAE**I**MRRT**G**-TV**I**T**V**DT
1ZZK (A:24-51)	d.51.1.1	LA**G**S**IIG**KG**G**QRIKQ**I**RHES**G**-AS**I**K**I**DE
2ASB (A:231-258)	d.52.3.1	QL**S**L**AIG**KE**G**QNARL**A**ARLT**G**-WR**I**D**I**RG
1K0R (A:231-259)	d.52.3.1	AK**G**A**CIG**PM**G**QRVRN**V**MSEL**S**GEK**I**D**I**ID

Table A.11: KH-motif (Fragment 11). Found in 2 folds comprising 2 superfamilies. This motif has been previously described [Grishin, 2001b]. We detected this fragment at HHsearch probabilities ranging between 80%-95%. The KH motif is involved in binding RNA or single-stranded DNA.

PDB accession	SCOP ID	Alignment
		hhhhhhhhhhhhhccceeeeeecceeeeec
1IN0 (A:106-134)	d.58.49.1	**M**AKK**I**T**KLV**KDSKIK**VQ**TQI**QG**EQ**VRV**T**G**
1WIH (A:46-74)	d.67.3.1	**C**TAA**A**I**KAI**RESGMN**LN**PEV**EG**TL**IRV**P**I**
1IS1 (A:74-102)	d.67.3.1	**L**TQK**V**E**KAI**MMSDLG**LN**PMS**AG**TI**IRV**P**L**

Table A.12: α-β-β-motif from ribosome recycling factor (Fragment 12). Found in 2 folds comprising 2 superfamilies. We detected this fragment at HHsearch probabilities ranging between 55%-70%.

PDB accession	SCOP ID	Alignment
		hhhhhhhhhhccbbbb
1JJ2 (K:127-142)	c.12.1.1	EG**AR**EK**VE**G**AGG**S**VE**L
2GYA (J:129-143)	c.12.1.1	KG**AR**AA**IE**A**AGG**K**IE**-
1CTF (A:104-119)	d.45.1.1	EA**LK**KA**LE**E**AGA**E**VE**V
1DD3 (A:112-127)	d.45.1.1	EE**IK**KK**LE**E**AGA**E**VE**L

Table A.13: α-β-motif from ribosomal proteins L15p and L18e (Fragment 13). Found in 2 folds comprising 2 superfamilies. We detected this fragment at HHsearch probabilities ranging between 50%-55%.

PDB accession	SCOP ID	Alignment
		ccceehhhhhhhcccccchhhhhhhh
1QW2 (A:88-113)	d.249.1.1	K**VV**EASQ**EAQKVGI**NPGDV**L**RN**V**IDK
1JX4 (A:42-67)	e.8.1.7	A**VA**TANY**EARKFGV**KAGIP**I**VE**A**KKI
1T94 (A:135-160)	e.8.1.7	M**LS**TSNY**HARRFGV**RAAMP**G**FI**A**KRL
1T3N (A:62-87)	e.8.1.7	L**VV**TCNY**EARKLGV**KKLMN**V**RD**A**KEK

Table A.14: DNA-binding β-α-α-motif (Fragment 14). Found in 2 folds comprising 2 superfamilies. We detected this fragment at HHsearch probabilities ranging between 80%-90%. This motif is involved in interactions with DNA (see the structure of 1JX4).

PDB accession	SCOP ID	Alignment
		eeeeceec-ccceeeccccceeee
1FM0 (D:53-74)	d.15.3.1	AAV**N**QT**LV**-SFDHP**L**TD**GD**E**V**A**F**
1RYJ (A:45-66)	d.15.3.2	VKK**N**GQ**IV**-IDEEE**I**FD**GD**I**I**E**V**
1TKE (A:38-59)	d.15.10.1	GRV**N**GE**LV**-DACDL**I**ENDAQ**L**S**I**
2FF4 (A:350-371)	b.26.1.2	VHV**Q**HER**I**-RSAVT**L**ND**GD**H**I**R**I**
2AFF (A:69-90)	b.26.1.2	TQV**N**GS**VI**-DEPVR**L**KH**GD**V**I**T**I**
1DM9 (A:36-57)	d.66.1.3	VHY**N**GQRS-KPSKI**V**ELNAT**L**T**L**
2GY9 (D:122-144)	d.66.1.2	IMV**N**GR**VV**NIASYQ**V**SPN**D**V**V**S**I**

Table A.15: RNA-binding α-L-motif (Fragment 15). Found in 3 folds comprising 4 superfamilies. We detected this fragment at HHsearch probabilities ranging between 60%-90%. This motif has been suggested to be involved in binding to RNA [Staker et al., 2000].

PDB accession	SCOP ID	Alignment
		eeeeeeccccceeeee
1MBY (A:861-875)	d.223.1	AVW**V**QFN**DGS**Q**LVM**Q
2F69 (A:118-132)	b.76.2	VCW**I**YYP**DGG**S**LVG**E
2GFA (A:985-999)	b.34.9	MYQ**V**EFE**DGS**Q**LVV**K

Table A.16: β-hairpin from tudor domain (Fragment 16). Found in 3 folds comprising 3 superfamilies. We detected this fragment at HHsearch probabilities ranging between 50%-55%.

PDB accession	SCOP ID	Alignment
		eeeechhhcccccceecccccceeeeee
1FEU (A:148-174)	b.53.1.1	DS**L**H**A**S**DL**KLPPG**VELA**VSPEET**IA**A**V**
1BDF (A:108-133)	d.181.1.1	GP**V**T**A**A**DI**THDGD**VEIV**-KPQHV**IC**H**L**

Table A.17: β-β-β-motif from ribosomal protein L25-like superfamily (Fragment 17). Found in 2 folds comprising 2 superfamilies. We detected this fragment at HHsearch probabilities ranging between 60%-65%.

PDB accession	SCOP ID	Alignment
		ccceeeeeccceeeeeeccccccceeeeeeeeeccc
1EX4 (A:238-264)	b.34.7.1	**PA**KL**LW**KGEGA**VVI**QD-----NSDIK**VVP**RRK
1KPF (A:28-59)	d.13.1.1	**PA**KI**IF**EDDRC**LAF**HDISPQAPTHFL**VIP**KKH
1FIT (A:12-43)	d.13.1.1	**PS**VV**FL**KTELS**FAL**VNRKPVVPGHVL**VCP**LRP
1Y23 (A:18-49)	d.13.1.1	**PS**AK**VY**EDEHV**LAF**LDISQVTKGHTL**VIP**KTH

Table A.18: β-meander from retroviral integrase (Fragment 18). Found in 2 folds comprising 2 superfamilies. We detected this fragment at HHsearch probabilities ranging between 50%-55%. This motif is implicated to be involved in binding DNA in retroviral integrase [Chen et al., 2000].

PDB accession	SCOP ID	Alignment
		eeeeeeeeeecccceeeeeecccchhhhhhhhhhhhhhhhhhhh
1UIL (A:57–100)	d.50.1.1	S**F**I**AEM**T**I**Y**I**KQL**G**RR**I**F**A**R**E**HGSNKK**LAA**QSC**A**LS**L**VR**QL**YH**L**
2YWQ (A:53–94)	d.204.1.1	K**A**R**AEI**Q**V**D**L**P−−**G**GL**V**R**V**E**E**EDADLY**AAI**DRA**V**DR**L**ET**QV**KR**F**

Table A.19: β-β-α-element from the double-stranded-RNA-binding domain (Fragment 19). Found in 2 folds comprising 2 superfamilies. We detected this fragment at HHsearch probabilities ranging between 50%-75%. This motif is involved in binding double-stranded RNA and in protein-protein interactions [Tian et al., 2004].

PDB accession	SCOP ID	Alignment
		hhhhhhhhhhhhh--chhhhhcchhhhhhhhh
2C9W (A:162-192)	a.271.1.1	**L**QHL**C**RLT**I**NK**C**T--GA**I**WG**L**P**L**PTR**L**K**D**Y**LE**
1LM8 (A:158-189)	b.3.3.1	**L**KER**C**LQV**V**RS**L**VKPEN**Y**RR**L**D**I**VRS**L**Y**E**D**LE**

Table A.20: SOCS Box motif (Fragment 20). Found in 2 folds comprising 2 superfamilies. We detected this fragment at HHsearch probabilities ranging between 80%-90%. This motif has been previously described [Zhang et al., 1999]. This motif is involved in interactions with elongins B and C. We note that this motif may represent a small, autonomous domain.

PDB accession	SCOP ID	Alignment
		ceeeeec-chhhhhhhhhhhh-hhcc-eeeeec
2NAC (A:192-221)	c.2.1.4	MH**VG**TV**A**-A**G**RI**G**LAV**L**RR**L**A-PFD**V**-HLH**Y**TD
1E5Q (A:4-33)	c.2.1.3	KS**VL**ML**G**-S**G**FVTRPT**L**DV**L**T-DS**GI**-KVT**VA**C
2BS2 (A:6-35)	c.3.1.4	CDS**L**VI**G**-G**G**LA**G**LRA**A**VATQ-QK**GL**-STI**VL**S
1GES (A:5-35)	c.3.1.5	YD**YI**AI**G**-G**G**SG**G**IAS**I**NR**A**A-MY**G**Q-KCA**LI**E
1DJQ (A:390-419)	c.4.1.1	DS**VL**IV**G**-A**G**PS**G**SEA**A**RV**L**M-ES**GY**-TVH**L**TD
1C0P (A:1005-1034)	c.4.1.2	KR**VV**VL**G**-S**G**VI**G**LSS**A**LI**L**A-RK**GY**-SVH**IL**A
2JFG (A:6-35)	c.5.1.1	KN**VV**II**G**-L**G**LT**G**LSC**V**DF**F**L-AR**GV**-TPR**VM**D
2J9G (A:3-32)	c.30.1.1	DK**IV**IAN-RGEI**A**LRI**L**RA**C**K-EL**GI**-KTV**AV**H
3ETJ (A:2-31)	c.30.1.1	KQ**VC**VL**G**-N**G**QL**G**RMLRQA**G**E-PL**GI**-AVWP**V**G
1RI5 (A:66-94)	c.66.1.34	DS**VL**DL**G**-C**G**K--GGD**L**LK**Y**E-RA**GI**gEYY**GV**D
1DUV (G:156-187)	c.78.1.1	MT**LV**YA**G**dARNNMGNS**M**LE**A**AaLT**GL**-DLR**LV**A
1V71 (A:75-104)	c.79.1.1	AG**VL**TF**S**-S**G**NH**A**QAI**A**LS**A**K-IL**GI**-PAK**II**M
1JW9 (A:32-62)	c.111.1.1	SR**VL**IV**G**-L**G**GL**G**CAASQY**L**A-SA**GV**gNLT**LL**D
1Q7E (A:10-39)	c.123.1.1	IK**VL**DFT-GVQS**G**PSCTQM**L**A-WF**GA**-DVIK**I**E

Table A.21: Dinucleotide-binding β-α-β motif (Fragment 21). Found in 10 folds comprising 10 superfamilies. We detected this fragment at HHsearch probabilities ranging between 50%-95%. This motif has been previously described [Bellamacina, 1996; Lupas et al., 2001]. This motif is found in different α/β Rossmann-like folds, in which it is involved in binding dinucleotides.

PDB accession	SCOP ID	Alignment
		ceeeeeccccchhhhhhhhhhhh
1ZNW (A:22-44)	c.37.1.1	R**VV**V**LSG**PSAV**GKSTVV**RC**L**RER
1RZ3 (A:23-45)	c.37.1.6	L**VL**G**I**D**G**LSRS**GKTTLA**NQ**L**SQT
2GC6 (A:40-60)	c.72.2.2	R**YI**H**VTG**TN--**GKGSAA**NA**I**AHV
2JFG (A:105-125)	c.72.2.1	P**IV**A**ITG**SN--**GKSTVT**TL**V**GEM
1KO7 (A:145-167)	c.91.1.2	V**GV**L**ITG**DSGI**GKS**E**TA**LE**L**IKR
1KNX (A:148-170)	c.91.1.2	V**GV**L**LTG**RSGI**GKS**E**CA**LD**L**INK

Table A.22: P-loop-motif (Fragment 22). Found in 3 folds comprising 3 superfamilies. This motif has been previously described [Gay and Walker, 1983; Matte et al., 1996; Lupas et al., 2001]. We detected this fragment at HHsearch probabilities ranging between 70%-95%. This motif functions by binding the phosphate backbone of a mononucleotide.

PDB accession	SCOP ID	Alignment
		hhhhhcccc
1I8O (A:10-18)	a.3.1.1	FKQ**C**MT**CH**R
1C52 (A:8-16)	a.3.1.1	YAQ**C**AG**CH**Q
1E85 (A:113-121)	a.24.3.2	GAS**C**KA**CH**D
1SP3 (A:19-27)	a.138.1.3	TTQ**C**LT**CH**E
1E2W (A:18-26)	b.2.6.1	RIV**C**AN**CH**L

Table A.23: Cytochrome-heme-attachment-motif (Fragment 23). Found in 2 folds comprising 2 superfamilies. This motif has been described previously [Lupas et al., 2001]. We detected this fragment at HHsearch probabilities ranging between 50%-65%. This motif binds covalently to a heme group.

PDB accession	SCOP ID	Alignment
		cccccchhhhhhhcc
2BS2 (B:149-163)*	a.1.2.1	DR**CI**EC**GCC**IA**AC**GT
2BS2 (B:206-220)*	a.1.2.1	FG**CM**TLL**AC**HD**VCP**K
1KF6 (B:146-160)*	a.1.2.1	SG**CI**NC**GLC**YA**ACP**Q
1KF6 (B:202-216)*	a.1.2.1	WS**C**TFV**GYC**SE**VCP**K
2FDN (A:6-20)*	d.58.1.1	EA**CI**SC**GAC**EPE**CP**V
2FDN (A:35-49)*	d.58.1.1	DT**CI**DC**GAC**AG**VCP**V

Table A.24: FeS-binding peptide (Fragment 24). Found in 2 folds comprising 2 superfamilies. We detected this fragment at HHsearch probabilities ranging between 80%-98%. This motif has been described previously [Lupas et al., 2001]. Both the alpha-helical ferredoxin superfamily (a.1.2) and the 4Fe-4S ferredoxins superfamily contain two homologous copies of this motif (indicated by a * in the alignment). The FeS-cluster is coordinated with three cysteines contributed by the first FeS-binding motif and one cysteine by the second motif.

PDB accession	SCOP ID	Alignment
		cccccccccchhhhhhhhhh
2SCP (B:104-123)*	a.39.1.5	**D**T**N**E**D**NN**I**SRD**EY**G**IF**FGM**L**
2SCP (B:138-157)*	a.39.1.5	**D**T**N**N**D**GL**L**SLE**EF**V**IA**GSD**F**
2CCL (B:2-21)*	a.139.1.1	**D**V**N**G**D**GT**I**NST**DL**T**ML**KRS**V**
2CCL (B:36-55)*	a.139.1.1	**D**V**D**K**N**GS**I**NAA**DV**L**LL**SRY**L**

Table A.25: Calcium-binding loop-helix-motif (Fragment 25). Found in 2 folds comprising 2 superfamilies. We detected this fragment at HHsearch probabilities ranging between 50%-65%. This motif binds calcium. Both the EF-hand superfamily and the type I dockerin superfamily contain two tandem copies of this motif (indicated by a * in the alignment).

PDB accession	SCOP ID	Alignment
		eecccccceechhhhhhhhhhhh
2GIV (A:35-57)	d.108.1.1	**W**L**C**EY**CLK**Y**M**KYEKS**Y**RF**HL**GQ**C**
1M36 (A:8-30)	g.37.1.2	**Y**L**C**EF**CLK**Y**M**KSRTI**L**QQ**HM**KK**C**
1X6H (A:48-70)	g.37.1.1	**F**V**C**SK**CGK**T**F**TRRNT**M**AR**HA**DN**C**
2DRP (A:141-163)	g.37.1.1	**Y**P**C**PF**CFK**E**F**TRKDN**M**TA**HV**KII

Table A.26: Zn-binding β-β-α-motif (Fragment 26). Found in 2 folds comprising 2 superfamilies. We detected this fragment at HHsearch probabilities ranging between 70%-80%. This motif is involved in coordinating Zn.

PDB accession	SCOP ID	Alignment
		ceecccceeccccccccccccchhhhhehhhhhhh
1VYX (A:7-40)	g.44.1.3	PV**C**WI**C**NEELGNERFR**A**CG**C**TGELEN**VH**RS**C**LST
1PTQ (A:242-275)	g.49.1.1	TF**C**DH**C**GSLLWGLVKQ**G**LK**C**EDCGMN**VH**HK**C**REK
1F62 (A:1-32)	g.50.1.2	AR**C**KV**C**RKK--GEDDK**L**IL**C**DECNKA**FH**LF**C**LRP

Table A.27: Zn-binding treble-clef-motif (Fragment 27). Found in 3 folds comprising 3 superfamilies. We detected this fragment at HHsearch probabilities ranging between 80%-95%. This motif has been described previously [Grishin, 2001c]. This motif coordinates two zinc ions.

PDB accession	SCOP ID	Alignment
		cccccccccccceeecccceeeeccccceeec
1L1O (C:479-508)	b.40.4.3	QA**CP**TQDCNKK**V**IDQQN**GLY**R**C**EK**C**DTE**F**P
1JJ2 (Y:37-65)	g.41.8.1	HK**CP**VC-GFKK**L**KRAGT**GIW**M**C**GH**C**GYK**I**A
2AKL (A:6-33)	g.41.3.5	PP**CP**--QCNSE**Y**TYEDG**ALL**V**C**PE**C**AHE**W**S

Table A.28: Zn-ribbon-motif (Fragment 28). Found in 2 folds comprising 3 superfamilies. This motif has been described previously [Grishin, 2000]. We detected this fragment at HHsearch probabilities ranging between 50%-70%. This motif coordinates a zinc ion.

PDB accession	SCOP ID	Alignment
		cccccccccccceee
1EP3 (B:223-237)	c.25.1.3	R**M**A**C**G**IGAC**YA**CV**EH
1QLA (B:54-68)	d.15.4.2	D**F**V**C**R**AGIC**GS**CG**MM
1CZP (A:38-52)	d.15.4.1	P**F**S**C**R**AGAC**ST**CA**GK

Table A.29: FeS-binding peptide (Fragment 29). Found in 2 folds comprising 2 superfamilies. We detected this fragment at HHsearch probabilities ranging between 75%-95%. This motif coordinates a 2Fe-2S cluster.

PDB accession	SCOP ID	Alignment
		eeeccccceeecccce-eec
1CCV (A:36-52)	g.22.1.1	**C**Q**C**QE**GF**LRNGEGA-**CV**L
1IJQ (A:675-692)	g.3.11.1	**C**A**C**PD**GM**LLARDMRS**CL**T
2BZ6 (L:112-129)	g.3.11.1	**C**R**C**HE**GY**SLLADGVS**C**TP
1D0G (R:26-43)	g.24.1.1	GL**C**PP**G**HHISEDGRD**CI**S
1EXT (A:137-152)	g.24.1.1	**C**T**C**HA**GF**FLR--ENE**CV**S

Table A.30: Cysteine-rich β-meander (Fragment 30). Found in 3 folds comprising 3 superfamilies. We detected this fragment at HHsearch probabilities ranging between 70%-90%.

PDB accession	SCOP ID	Alignment
		hhhhhhhhhhhhhhcchhhhhhhhhhhhhh
2E2A (A:19-50)	a.7.2.1	DARSKLLE**A**LK**A**AEN**GD**FAK**A**DSL**V**VE**A**GSCI
1WR0 (A:14-45)	a.7.14.1	KAID**L**ASK**A**AQEDKA**GN**YEE**A**LQL**Y**QH**A**VQYF
2CRB (A:14-45)	a.7.16.1	LAHQQSRR**A**DR**L**LAA**G**KYEE**A**ISCHRK**A**TTYL
2PMR (A:36-67)	a.8.11.1	RALN**Y**RDD**S**VY**Y**LEK**GD**HIT**S**FGC**I**TY**A**HGLL
1OM2 (A:19-50)	a.23.4.1	FFLEEIQL**G**EE**L**LAQ**GD**YEK**G**VDH**L**TN**A**IAVC
1O3U (A:3-34)	a.24.16.3	AAKDDLEH**A**KHDLEH**G**FYNW**A**CFSSQQ**A**AEKA
1UG7 (A:18-49)	a.24.24.1	RWGASLRR**G**AD**F**DSW**GQ**LVE**A**IDE**Y**QI**L**ARHL
1QSA (A:412-443)	a.118.5.1	SKTEQAQL**A**RY**A**FNNQWWDL**S**VQATIA**G**Klwd
2O02 (A:2-33)	a.118.7.1	DKNE**L**VQK**A**KL**A**EQAERYDD**M**AAC**M**KS**V**TEQG
1ELW (A:3-34)*	a.118.8.1	QVNE**L**KEK**G**NK**A**LSV**GN**IDD**A**LQC**Y**SE**A**IKLD
1ELW (A:37-68)*	a.118.8.1	NHVL**Y**SNR**S**AA**Y**AKK**GD**YQK**A**YED**G**CKTVDLK
1ELW (A:71-102)*	a.118.8.1	WGKG**Y**SRK**A**AA**L**EFLNRFEE**A**KRT**Y**EE**G**LKHE
2D2S (A:535-566)	a.118.17.2	FLDE**G**VEE**I**DIELARLRFES**A**VET**L**LD**I**ESQL
1WY6 (A:125-156)	a.118.20.1	SASI**L**VAI**A**NA**L**RRV**GD**ERD**A**TTL**L**IE**A**CKKG
2A9U (A:39-70)	a.118.23.1	SALK**I**FKT**A**EE**C**RLDR**D**EER**A**YVL**Y**MK**Y**VTVY
2IJQ (A:29-60)	a.246.2.1	TLRR**A**VVH**G**VR**L**YNS**GE**FHE**S**HDC**F**EDEWYNY
2CFU (A:448-479)	d.157.1.13	GAER**L**LEQ**A**RASYAR**GE**YRW**V**VEV**V**NR**L**VFAE
1ZBP (A:30-61)	e.61.1.1	DASLRSSF**I**EL**L**CID**GD**FER**A**DEQ**L**MQ**S**IKLF

Table A.31: Tetratricopeptide repeat (TPR) element (Fragment 31). Found in 8 folds comprising 16 superfamilies. This motif has been previously described [Goebl and Yanagida, 1991; D'Andrea and Regan, 2003]. We detected this fragment at HHsearch probabilities ranging between 55%-95%. The TPR element mediates protein-protein interactions and the assembly of multiprotein complexes. Sequence alignment of TPR elements does not show many conserved columns of identical residues, but it shows columns of small and large amino acids. The TPR-like superfamily (a.118.8) contains multiple tandem copies of this motif (indicated by a * in the alignment).

PDB accession	SCOP ID	Alignment
		hhhhhhhhhhhhhhcccchhhhhhhhh
2QDY (B:32-59)*	b.34.4.4	**W**EHLPYS**L**MF**AG**VAELG**AF**SVDE**V**RY**V**V
2QDY (B:73-99)*	b.34.4.4	**Y**ERYVIG**V**AT**LM**VE-KG**IL**TQDE**L**ES**L**A
2QDY (A:19-45)	d.149.1.1	**V**SDRAWA**L**FR**AL**DG-KG**LV**PDGY**V**EG**W**K
1V29 (A:29-55)	d.149.1.1	**W**EARAKA**L**ES**LL**IE-KR**LL**SSDA**I**ER**V**I

Table A.32: α-α-motif from nitrile hydratase (Fragment 32). Found in 2 folds comprising 2 superfamilies. We detected this fragment at HHsearch probabilities ranging between 70%-80%. Nitrile hydratases are composed of two subunits, termed α and β. The N-terminal subdomains of these subunits share a homologous α-α-motif. The β chain (b.34.4) contains two tandem copies of this motif (indicated by a * in the alignment).

PDB accession	SCOP ID	Alignment
		hhhhhhhhhhhhhhhhhhhhhhhhhhhhhhcchhhhhhhhhhhhhhhhhhhhhhhhhhhhh
1EK9 (A:117-177)*	f.5.1.1	LNAIDV**L**SYTQAQKEAIYRQ**L**DQTTQR**F**NVGLVAIT**DV**Q**NA**RAQ**Y**DTVLANELTA**R**N**NL**DN
1EK9 (A:335-395)*	f.5.1.1	NASISS**I**NAYKQAVVSAQSS**L**DAMEAG**Y**SVGTRTIV**DV**L**DA**TTT**L**YNAKQELANA**R**Y**NY**LI
1VF7 (A:74-134)	f.46.1.1	ATYEAD**Y**QSAQANLASTQEQ**A**QRYKLL**V**ADQAVSKQ**QY**A**DA**NAA**Y**LQSKAAVEQA**R**I**NL**RY

Table A.33: α-α-motif from TolC (Fragment 33). Found in 2 folds comprising 2 superfamilies. We detected this fragment at HHsearch probabilities ranging between 50%-65%. Outer membrane efflux proteins (f.5.1) contain two tandem copies of this motif (indicated by a * in the alignment).

PDB accession	SCOP ID	Alignment
		hhhhhhhhhhhhhhhhhhhhhhcchhhhhhhhhhhhhhhhhhhhh
2NW8 (A:56–102)	a.266.1.1	QT**S**E**LWL**K**LLA**HE**LRAAI**VH**L**QR**DEV**WQCRKV**LAR**SKQ**VL**RQ**L**TEQ**W**
1VH6 (A:22–68)	a.24.19.1	EL**T**L**MLY**N**GCL**KF**IRLAA**QA**I**EN**DDM**ERKNEN**LIK**AQN**II**QE**L**NFT**L**

Table A.34: α-α-motif from flagellar export chaperone FliS (Fragment 34). Found in 2 folds comprising 2 superfamilies. We detected this fragment at HHsearch probabilities ranging between 60%-70%.

PDB accession	SCOP ID	Alignment
		ceeeeeeccccceeeeeeecc--eeeeeecccc-----ceeeeeec
1GEN (A:567-604)	b.66.1.1	R**V**D**AA**FNWSKNKK**TY**IFAG**D**--**KF**WR**YN**EVKK-----KMDPGFPK
1TL2 (A:42-83)	b.67.1.1	N**F**K**FL**F**L**SPGG-E**LY**GVLN**D**--**KI**YK**G**TPPTHDNDNWMGRAKKIG
1CRU (A:28-64)	b.68.2.1	KPH**AL**L**W**GPD-NQ**IW**LTERATG**KI**LR**VN**PES------GSV-KTVF
1TBG (A:57-94)*	b.69.4.1	K**I**Y**AM**H**W**GTDSRL**LV**SASQ**D**-G**KL**II**WD**SYT------TNKVHAIP
1TBG (A:272-309)*	b.69.4.1	G**I**TS**V**S**F**SKSGRL**LL**AGYD**D**-FNCNV**WD**ALK------ADRAGVLA
1HZU (A:180-219)	b.70.2.1	A**V**H**I**SR**M**SASGRY**LL**VIGR**D**-A**RI**DM**ID**LWAKE----PTKVAEIK

Table A.35: Four-stranded β-meander from β-propellers (Fragment 35). Found in 5 folds comprising 30 superfamilies. This fragment has been described previously [Chaudhuri et al., 2008]. We detected this fragment at HHsearch probabilities ranging between 50%-95%. β-propellers form a circular fold comprising between four and ten homologous copies of this motif. Two repeats from the seven-bladed β-propeller 1TBG are shown (indicated by a * in the alignment).

PDB accession	SCOP ID	Alignment
		eeeeeccccceeeee
1QBA (A:842-855)	b.1.18.2	IEY**S**T**D**G**G**KQ**W**QRY
1D7P (M:2262-2274)	b.18.1.2	ISS**S**Q**D**-**G**HQ**W**TLF
1VKD (A:73-85)*	b.67.2.4	FGR**S**K**D**-**G**IN**W**EIE
1VKD (A:131-143)*	b.67.2.4	VGM**T**K**D**-**F**KT**F**VRL
1VKD (B:184-196)*	b.67.2.4	LSE**S**P**D**-**M**IH**W**GNH
3SIL (A:70-83)*	b.68.1.1	AAR**S**T**D**G**G**KT**W**NKK
3SIL (A:253-266)*	b.68.1.1	SFE**T**K**D**F**G**KT**W**TEF
1SQJ (A:194-207)	b.69.13.1	MYV**T**H**D**G**G**VS**W**EPV
1LNI (A:79-92)	d.1.1.2	DYY**T**G**D**H**Y**AT**F**SLI

Table A.36: ASP-box (Fragment 36). Found in 6 folds comprising 6 superfamilies. This motif has been previously described [Copley et al., 2001]. We detected this fragment at HHsearch probabilities ranging between 50%-90%. This motif has a characteristic SxDxxxW sequence motif. Most ASP-box-containing domains act on carbohydrates; a direct role for this motif in carbohydrate binding remains, however, unclear. Some folds contain multiple copies of this motif (indicated by a * in the alignment).

PDB accession	SCOP ID	Alignment
		bbbbccccbbbcccbbbb
1BDO (A:100-117)*	b.84.1.1	KAF**I**EV**GQ**K**V**NV**GD**T**LC**I
1BDO (A:137-154)*	b.84.1.1	AIL**V**ES**GQ**P**V**EFD**E**P**LV**V
1E2W (A:213-230)	b.84.2.2	DLI**V**KE**GQ**T**V**QAD**Q**P**L**TN
1GPR (A:97-114)	b.84.3.1	TSF**V**SE**GD**R**V**EP**GQ**K**LL**E
1V8Q (A:22-39)	b.84.4.1	GVKRYE**GQ**V**V**RA**GN**I**LV**R
1QPO (A:75-92)	d.41.2.1	LDR**V**ED**G**AR**V**PP**GE**A**LM**T
1BRW (A:379-396)	d.41.3.1	VLHKKI**GD**R**V**QK**GE**A**LA**T
1SMY (C:640-657)	e.29.1.1	RPR**V**VV**GQ**R**V**RK**GD**L**LA**D
1VF7 (A:53-70)	f.46.1.1	KRL**F**KE**G**SD**V**KA**GQ**Q**LY**Q

Table A.37: β-β-β-hammerhead-motif (Fragment 37). Found in 4 folds comprising 8 superfamilies. We detected this fragment at HHsearch probabilities ranging between 80%-95%. The barrel-sandwich hybrid fold (b.84.1) is pseudo-symmetric and contains two homologous copies of this motif.

PDB accession	SCOP ID	Alignment
		bbbbcccccbbbbbb
1LCY (A:175-189)	b.47.1.1	G**PLV**NLD**G**E**VIGV**NT
1GDN (A:197-211)	b.47.1.2	G**PIV**DSSNT**LIGA**VS
1TIF (A:16-30)	d.15.8.1	VR**LI**DQN**G**DQ**LGI**KS
2NYC (A:223-237)*	d.37.1.1	FF**VV**DDV**G**R**LVGV**LT
2NYC (A:295-309)*	d.37.1.1	V**PII**DEN**G**Y**LI**N**V**YE
1F5M (A:135-149)	d.110.2.1	V**PII**SND**G**KT**LGV**ID
1N9L (A:104-118)	d.110.3.6	T**PI**KTPD**G**R**V**SK**F**VG
1P0Z (A:110-124)	d.110.6.1	S**PI**QDAT**G**K**VIGI**VS
1U7Q (A:430-444)	d.220.1.1	PL**VV**VEGSR**VLGV**IA

Table A.38: β-hairpin from CBS-domain (Fragment 38). Found in 5 folds comprising 7 superfamilies. We detected this fragment at HHsearch probabilities ranging between 50%-60%. In trypsin-like serine proteases (b.47.1) this motif contains residues important for intermolecular hydrophobic contacts with PDZ domain and the catalytic serine lies at the N-terminal end of the first β-strand. The CBS-domain (d.37.1) contains two copies of this motif (indicated by a * in the alignment).

PDB accession	SCOP ID	Alignment
		eeeeeeeee-cceeeeeeee
1QJP (A:124-142)	f.4.1.1	A**GGV**E**Y**A**I**T-PEIATR**L**E**Y**Q
1QD6 (A:196-213)	f.4.2.1	Q**L**K**I**G**Y**H**L**G--DAV**L**S**A**K**G**Q
1T16 (A:81-99)*	f.4.3.4	N**M**H**F**V**A**P**I**N-DQFG**W**G**A**S**I**T
1T16 (A:129-147)*	f.4.3.4	N**LSG**A**Y**R**L**N-NAWS**F**G**L**G**F**N
1T16 (A:210-228)*	f.4.3.4	N**AGI**L**Y**E**L**D-KNNR**Y**A**L**T**Y**R
1T16 (A:276-294)*	f.4.3.4	E**VSG**YNR**V**D-PQWA**I**H**Y**S**L**A
1T16 (A:328-346)*	f.4.3.4	A**LG**TT**Y**Y**Y**D-DNWT**F**RTG**I**A
1T16 (A:370-388)*	f.4.3.4	S**AG**TT**Y**A**F**N-KDAS**V**D**V**G**V**S
1I78 (A:234-252)	f.4.4.1	A**V**N**A**G**Y**Y**V**T-PNAK**V**Y**V**E**G**A
1UYN (A:886-904)	f.4.5.1	F**AGI**RHD**A**G-DIGY**L**K**G**L**F**S
1TLY (A:217-234)	f.4.6.1	SH**IL**A**L**N**Y**--DHWH**Y**S**V**V**A**R
2GR8 (A:1062-1081)	d.24.1.4	A**IGV**SRISDNGKVI**I**R**L**S**G**T

Table A.39: β-hairpin from outer membrane β-barrel (Fragment 39). Found in 2 folds comprising 7 superfamilies. This motif has been described previously [Remmert et al., 2009]. We detected this fragment at HHsearch probabilities ranging between 50%-70%. Each of these folds contain multiple copies of this fragment.

PDB accession	SCOP ID	Alignment
		ceeeeeecceeeeeeec
1YRR (A:18-34)	b.92.1.5	**D**H**AVVI**A**D**GL**I**KS**VC**P**V**
2OOD (A:35-51)	b.92.1.4	**D**G**LMVV**T**D**GV**I**KA**FG**P**Y**
2FB5 (A:135-151)	d.320.1.1	**D**G**AVLV**K**N**NH**I**VS**AA**N**I**

Table A.40: β-hairpin from the composite domain of metallo-dependent hydrolases (Fragment 40). Found in 2 folds comprising 2 superfamilies. We detected this fragment at HHsearch probabilities ranging between 60%-65%.

PDB accession	SCOP ID	Alignment
		eeeechhhhhhcccccccceeeee
1CZ5 (A:23-45)*	b.52.2.3	RVR**L**DESSRRL**L**D**A**EI**G**DV**V**E**I**E
1CZ5 (A:66-88)*	b.52.2.3	IVR**I**DSV**M**RNN**C**G**A**SI**GD**K**V**K**V**R
1YFB (A:17-39)	b.129.1.3	RVV**I**PIE**L**RRT**L**G**I**AEK**D**A**L**E**I**Y
2D9R (A:78-100)	b.129.2.1	ILG**L**RQD**I**RRA**I**GKQP**GD**S**V**Y**V**T
1A8P (A:19-41)	b.43.4.2	LFS**F**KTTRNPS**L**R**F**EN**GQ**F**V**M**I**G
2VBU (A:107-129)	b.43.5.2	EII**A**PMK**L**REQ**F**N**L**KD**GD**V**I**K**I**L
1YLE (A:314-336)	d.108.1.8	PVA**L**SVE**A**AEA**L**G**V**GE**G**AS**V**R**L**V
1RQP (A:217-239)	b.141.1.1	WTN**I**HRTDLEK**A**G**I**GY**G**AR**L**R**L**T
1WID (A:253-275)	b.142.1.2	LTK**G**WSR**F**VKEKN**L**RA**GD**V**V**S**F**S

Table A.41: GD-box-containing β-α-β-motif. Found in 6 folds comprising 8 superfamilies. This motif has been described previously by us [Coles et al., 1999, 2005, 2006; Alva et al., 2008]. We detected this fragment at HHsearch probabilities ranging between 60%-95%. These folds comprise two copies of this motif, either as monomers with internal sequence symmetry or as homodimers with one copy per subunit.

PDB accession	SCOP ID	Alignment
		bbbbbbcccccchhhhhhhhhhhhh
1D1Q (A:7-31)	c.44.1.1	IS**VAF**IALGNFC**RS**PM**A**EA**IF**KHEV
1VKR (A:378-402)	c.44.2.1	RK**IIV**A**C**DA**G**MGS**S**AM**G**AG**VL**RKKI
1VHR (A:118-142)	c.45.1.1	GR**VLV**H**C**RE**G**YS**RS**PT**L**VI**AY**LMMR
1D5R (A:118-142)	c.45.1.1	HV**AAI**H**C**KA**G**KG**RT**GV**M**IC**AY**LLHR
1YMK (A:467-491)	c.46.1.1	VI**LIF**H**C**EF**S**SE**R**GPR**M**CR**FI**RERD

Table A.42: β-α-motif from protein phosphatases (Fragment 42). Found in 3 folds comprising 4 superfamilies. This motif has been previously described [Lupas et al., 2001]. We detected this fragment at HHsearch probabilities ranging between 50%-80%. In these fold, the catalytically important cysteine and arginine residues lie within this motif.

PDB accession	SCOP ID	Alignment
		hhhhhhhhhhhhhhhhcccchhhceeeeee
1NBW (A:43-72)	c.55.1.6	RDNIAGT**L**AA**L**EQ**AL**AKTP**W**SMS**DV**SRIY**L**
1HUX (A:35-64)	c.55.1.5	GTGTSGP**A**RS**I**SE**VL**ENAH**M**KKE**DM**AFTL**A**
1OX0 (A:272-302)	c.95.1.1	HPEGQGA**I**KA**I**KL**AL**EEAE**I**SPE**QV**AYVN**A**
1TED (A:282-312)	c.95.1.2	GYIFSGV**A**PV**V**TE**ML**WDNG**L**QIS**DI**DLWA**I**

Table A.43: α-α-β motif from actin (Fragment 43). Found in 2 folds comprising 2 superfamilies. We detected this fragment at HHsearch probabilities ranging between 60%-80%.

PDB accession	SCOP ID	Alignment
		ceeeecccchhhhhhhhhhhcceeeeec
1NMO (A:188-213)	c.135.1.1	**DAFI**TGE**V**SEQT**I**HS**A**REQ**GL**HF**Y**AA
2GX8 (A:306-331)	c.135.1.1	**DVYV**TGD**M**YYHV**A**HD**A**MML**GL**NI**V**DP
1O13 (A:66-91)	c.55.5.1	**ELVI**VRG**I**GRRA**I**AA**F**EAM**GV**KV**I**KG
1P90 (A:167-192)	c.55.5.2	**QVLY**VVS**I**GGPA**A**AK**V**VRA**GI**HP**L**KK

Table A.44: β-α-β-motif from NIF3-like superfamily (Fragment 44). Found in 2 folds comprising 2 superfamilies. We detected this fragment at HHsearch probabilities ranging between 55%-70%.

PDB accession	SCOP ID	Alignment
		bbbbbcchhhhhhhh
1MLA (A:87-101)	c.19.1.1	**MMAG**H**S**L**G**EYSA**L**VC
1OXW (A:72-86)	c.19.1.3	**VIGG**T**S**T**GG**LLT**A**MI
1JJF (A:167-181)	c.69.1.2	**AIAG**L**S**M**GG**GQS**F**NI
3C8D (A:276-290)	c.69.1.2	**VVAG**Q**S**F**GG**LSA**L**YA
2PBL (A:131-145)	c.69.1.2	**VLAG**H**S**A**GG**HLV**A**RM
1MTZ (A:100-114)	c.69.1.7	**FLMG**S**S**Y**GG**ALA**L**AY

Table A.45: Serine nucleophilic elbow (Fragment 45). Found in 2 folds comprising 2 superfamilies. We detected this fragment at HHsearch probabilities ranging between 50%-65%. α/β-hydrolases (c.69.1) contain a classical ser-his-asp catalytic triad. The serine nucleophile is located at the tight turn of this motif.

PDB accession	SCOP ID	Alignment
		hhhhhhhhccceeeeeecc
1ECS (A:16-33)	d.32.1.1	STAA**FYE**RL**GF**GI**VF**RDA
1SS4 (A:21-38)	d.32.1.6	NAIS**FFE**EI**GL**NLE**G**RAN
1GHE (A:136-153)	d.108.1.1	VAEA**FY**SAL**AY**TR**VG**ELP
1N71 (A:142-159)	d.108.1.1	HPYE**FYE**KL**GY**KI**VG**VLP
1VHS (A:126-143)	d.108.1.1	PSLK**LFE**KH**GF**AE**WG**LFP

Table A.46: α-β-motif from Nat (Fragment 46). Found in 2 folds comprising 2 superfamilies. We detected this fragment at HHsearch probabilities ranging between 75%-95%. In Acyl-CoA N-acyltransferases (Nat; d.108.1) this motif is involved in binding Acyl-CoA.

PDB accession	SCOP ID	Alignment
		eeeeeecchhhhhhhhhhhh
1A9X (A:382-401)	d.142.1.2	**G**E**VMA**I**G**RTQQESLQK**A**LR**G**
1W96 (A:507-526)	b.84.2.1	**G**H**IFA**F**G**ENRQASRKH**M**VV**A**
1ULZ (A:386-405)	b.84.2.1	**A**K**LI**TW**A**PTWDEAVER**M**RA**A**
1KJQ (A:365-384)	b.84.2.1	**G**V**ALA**T**A**ESVVDAIER**A**KH**A**

Table A.47: β-α-motif from ATP-grasp fold (Fragment 47). Found in 2 folds comprising 2 superfamilies. We detected this fragment at HHsearch probabilities ranging between 85%-95%.

PDB accession	SCOP ID	Alignment
		eeeecccchhhhhhhhhhhhhccccceeeeecc
2AMH (A:11-40)	c.51.4.2	TM**IIGT**S**S**AF**R**ANV**L**REHFGDRFRN**F**V**L**LP
1EX2 (A:4-30)	c.51.4.2	PL**ILAS**Q**S**PR**R**KEL**L**DLLQL----P**Y**S**I**IV
1PDA (A:)	c.94.1.1	GS**IVGT**S**S**LR**R**QCQ**L**AERRP----D**L**I**I**RS

Table A.48: β-α-β-motif from ITPase-like superfamily (Fragment 48). Found in 2 folds comprising 2 superfamilies. We detected this fragment at HHsearch probabilities ranging between 50%-80%.

PDB accession	SCOP ID	Alignment
		cccccccchhhhhhhhhhh
1LBU (A:36-54)*	a.20.1.1	**L**AI**DG**Q**FG**PA**T**KAA**V**QR**F**Q
1LBU (A:59-77)*	a.20.1.1	**L**AA**DG**I**AG**PA**T**FNK**I**YQ**L**Q
2IKB (A:103-121)	d.2.1.9	**V**PD**DG**V**IG**AV**S**LKA**I**NS**L**P
2NR7 (A:112-130)	d.2.1.9	**V**QA**DG**I**VG**NK**T**LQA**V**NS**A**D

Table A.49: Loop-helix-motif from PGBD-like fold (Fragment 49). Found in 2 folds comprising 2 superfamilies. We detected this fragment at HHsearch probabilities ranging between 80%-95%. We found two homologous copies of this motif in the PGBD-like superfamily (a.20.1; indicated by a * in the alignment).

PDB accession	SCOP ID	Alignment
		ccchhhhhhhhhhhh
1E29 (A:106-120)	a.3.1.1	N**Y**TE**DDI**FD**V**AGYTL
1CC5 (A:71-85)	a.3.1.1	D**C**SD**DEL**KA**A**IGK**M**S
1KV9 (A:639-653)	a.3.1.6	S**L**KP**EEV**EQ**I**KLY**V**M
1W7C (A:63-77)	d.17.2.1	S**L**AK**EEV**QE**V**LDL**L**H
2OQE (A:27-41)	d.17.2.1	P**L**STA**EI**KA**A**TNT**V**K
1W2Z (A:12-26)	d.17.2.1	P**L**TK**EEF**LA**V**QTI**V**Q

Table A.50: Loop-helix-motif from Amine oxidase N-terminal superfamily (Fragment 50). Found in 2 folds comprising 2 superfamilies. We detected this fragment at HHsearch probabilities ranging between 60%-75%.

Appendix B

Contributions

Chapter 3

Section 3.2
This work is part of a manuscript that has been previously published in *Protein Science* [Alva et al., 2010]. Text and figures from this manuscript have been reproduced with permission from the publisher. This project was carried out in collaboration with Johannes Söding (JS), Michael Remmert (MR), Andreas Biegert (AB), and Andrei N. Lupas (ANL). I, JS and ANL conceived the study. JS produced the first version of *the galaxy of folds*. MR and AB prepared the data and produced the final version of the *the galaxy of folds*. I analyzed the data. I and ANL wrote the paper.

Section 3.3
I, Johannes Söding (JS), and Andrei N. Lupas (ANL) conceived the study. I performed the experiments and analyzed the data.

Chapter 4

This work is part of a manuscript that has been previously published in *BMC Structural Biology* [Alva et al., 2007]. Text and figures from this manuscript have been reproduced with permission from the publisher. I, Moritz Ammelburg (MA), and Andrei N. Lupas (ANL) conceived this study. Johannes Söding discovered the similarity between histones and C-domains. I and MA executed the analysis. I and ANL wrote the paper.

Chapter 5

This work is part of a manuscript that has been previously published in *Protein Science* [Alva et al., 2008]. Text and figures from this manuscript have been reproduced with permission from the publisher. I and Andrei N. Lupas (ANL) conceived this study. I performed the experiments and analyzed the data. Michael Habeck, Murray Coles, and Stanislaw Horkawicz-Dunnin contributed to the analysis. I and ANL wrote the paper.

Chapter 6

This work is part of a manuscript that has been previously published in *Current Opinion in Structural Biology* [Alva et al., 2008]. Text and figures from this manuscript have been reproduced with permission from the publisher. I, Murray Coles (MC), Kristin K. Koretke (KKK), and Andrei N. Lupas (ANL) conceived this study. I and ANL analyzed the data. MC and KKK contributed to the analysis. I, MC, and ANL wrote the paper.

Appendix C

Publications

- Bonitz T, **Alva V**, Saleh O, Lupas AN, Heide L. Evolutionary relationships of microbial aromatic prenyltransferases. PLoS One. 2011;6(11):e27336.

- Kopec KO, **Alva V**, Lupas AN. Bioinformatics of the TULIP domain superfamily. Biochem Soc Trans. 2011 Aug;39(4):1033-8.

- Kopec KO, **Alva V**, Lupas AN. Homology of SMP domains to the TULIP superfamily of lipid-binding proteins provides a structural basis for lipid exchange between ER and mitochondria. Bioinformatics. 2010 Aug 15;26(16):1927-31.

- **Alva V**, Remmert M, Biegert A, Lupas AN, Sding J. A galaxy of folds. Protein Sci. 2010 Jan;19(1):124-30.

  – Text and figures from this manuscript appear in Chapter 3.

- **Alva V**, Dunin-Horkawicz S, Habeck M, Coles M, Lupas AN. The GD box: a widespread noncontiguous supersecondary structural element. Protein Sci. 2009 Sep;18(9):1961-6.

  – Text and figures from this manuscript appear in Chapter 5.

- **Alva V**, Koretke KK, Coles M, Lupas AN. Cradle-loop barrels and the concept of metafolds in protein classification by natural descent. Curr Opin Struct Biol. 2008 Jun;18(3):358-65.

  – Text and figures from this manuscript appear in Chapter 6.

- **Alva V**, Syamala Devi DP, Sowdhamini R. COILCHECK: an interactive server for the analysis of interface regions in coiled coils. Protein Pept Lett. 2008;15(1):33-8.

- **Alva V**, Ammelburg M, Söding J, Lupas AN. On the origin of the histone fold. BMC Struct Biol. 2007 Mar 28;7:17.
  - Text and figures from this manuscript appear in Chapter 4.

- Ammelburg M, Hartmann MD, Djuranovic S, **Alva V**, Koretke KK, Martin J, Sauer G, Truffault V, Zeth K, Lupas AN, Coles M. A CTP-dependent archaeal riboflavin kinase forms a bridge in the evolution of cradle-loop barrels. Structure. 2007 Dec;15(12):1577-90.

Bibliography

Alber, F., Dokudovskaya, S., Veenhoff, L. M., Zhang, W., Kipper, J., Devos, D., Suprapto, A., Karni-Schmidt, O., Williams, R., Chait, B. T., Sali, A., and Rout, M. P. (2007). The molecular architecture of the nuclear pore complex. *Nature*, 450(7170):695–701.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *J Mol Biol*, 215(3):403–410.

Altschul, S. F., Madden, T. L., Schffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. (1997). Gapped blast and psi-blast: a new generation of protein database search programs. *Nucleic Acids Res*, 25(17):3389–3402.

Alva, V., Ammelburg, M., Söding, J., and Lupas, A. N. (2007). On the origin of the histone fold. *BMC Struct Biol*, 7:17–17.

Alva, V., Dunin-Horkawicz, S., Habeck, M., Coles, M., and Lupas, A. N. (2009). The gd box: a widespread noncontiguous supersecondary structural element. *Protein Sci*, 18(9):1961–1966.

Alva, V., Koretke, K. K., Coles, M., and Lupas, A. N. (2008). Cradle-loop barrels and the concept of metafolds in protein classification by natural descent. *Curr Opin Struct Biol*, 18(3):358–365.

Alva, V., Remmert, M., Biegert, A., Lupas, A. N., and Soeding, J. (2010). A galaxy of folds. *Protein Sci*, 19(1):124–130.

Ammelburg, M., Frickey, T., and Lupas, A. N. (2006). Classification of aaa+ proteins. *J Struct Biol*, 156(1):2–11.

Ammelburg, M., Hartmann, M. D., Djuranovic, S., Alva, V., Koretke, K. K., Martin, J., Sauer, G., Truffault, V., Zeth, K., Lupas, A. N., and Coles, M. (2007). A ctp-dependent archaeal riboflavin kinase forms a bridge in the evolution of cradle-loop barrels. *Structure*, 15(12):1577–1590.

Andreeva, A., Howorth, D., Brenner, S. E., Hubbard, T. J. P., Chothia, C., and Murzin, A. G. (2004). Scop database in 2004: refinements integrate structure and sequence family data. *Nucleic Acids Res*, 32(Database issue):D226–D229.

Andreeva, A., Howorth, D., Chandonia, J.-M., Brenner, S. E., Hubbard, T. J. P., Chothia, C., and Murzin, A. G. (2008). Data growth and its impact on the scop database: new developments. *Nucleic Acids Res*, 36(Database issue):D419–D425.

Andreeva, A. and Murzin, A. G. (2006). Evolution of protein fold in the presence of functional constraints. *Curr Opin Struct Biol*, 16(3):399–408.

Andreeva, A., Prlic, A., Hubbard, T. J. P., and Murzin, A. G. (2007). Sisyphus–structural alignments for proteins with non-trivial relationships. *Nucleic Acids Res*, 35(Database issue):D253–D259.

Anfinsen, C. B. (1973). Principles that govern the folding of protein chains. *Science*, 181(4096):223–230.

Aravind, L., Anantharaman, V., Balaji, S., Babu, M. M., and Iyer, L. M. (2005). The many faces of the helix-turn-helix domain: transcription regulation and beyond. *FEMS Microbiol Rev*, 29(2):231–262.

Arents, G., Burlingame, R. W., Wang, B. C., Love, W. E., and Moudrianakis, E. N. (1991). The nucleosomal core histone octamer at 3.1 A resolution: a tripartite protein assembly and a left-handed superhelix. *Proc Natl Acad Sci U S A*, 88:10148–10152.

Arents, G. and Moudrianakis, E. N. (1993). Topography of the histone octamer surface: Repeating structural motifs utilized in the docking of nucleosomal DNA. *Proc Natl Acad Sci U S A*, 90:10489–10493.

Arents, G. and Moudrianakis, E. N. (1995). The histone fold: a ubiquitous architectural motif utilized in DNA compaction and protein dimerization. *Proc Natl Acad Sci U S A*, 92:11170–11174.

Arnold, T., Poynor, M., Nussberger, S., Lupas, A. N., and Linke, D. (2007). Gene duplication of the eight-stranded beta-barrel ompx produces a functional pore: a scenario for the evolution of transmembrane beta-barrels. *J Mol Biol*, 366(4):1174–1184.

Bailey, K. A., Chow, C. S., and Reeve, J. N. (1999). Histone stoichiometry and DNA circularization in archaeal nucleosomes. *Nucleic Acids Res*, 27:532–536.

Ban, N., Nissen, P., Hansen, J., Moore, P. B., and Steitz, T. A. (2000). The complete atomic structure of the large ribosomal subunit at 2.4 a resolution. *Science*, 289(5481):905–920.

Baxevanis, A. D., Arents, G., Moudrianakis, E. N., and Landsman, D. (1995). A variety of DNA-binding and multimeric proteins contain the histone fold motif. *Nucleic Acids Res*, 23:2685–2691.

Bellamacina, C. R. (1996). The nicotinamide dinucleotide binding motif: a comparison of nucleotide binding proteins. *FASEB J*, 10(11):1257–1269.

Belogurov, G. A., Vassylyeva, M. N., Svetlov, V., Klyuyev, S., Grishin, N. V., Vassylyev, D. G., and Artsimovitch, I. (2007). Structural basis for converting a general transcription factor into an operon-specific virulence regulator. *Mol Cell*, 26(1):117–129.

Bennett, M. J., Schlunegger, M. P., and Eisenberg, D. (1995). 3d domain swapping: a mechanism for oligomer assembly. *Protein Sci*, 4(12):2455–2468.

Bharat, T. A. M., Eisenbeis, S., Zeth, K., and Hoecker, B. (2008). A beta alpha-barrel built by the combination of fragments from different folds. *Proc Natl Acad Sci U S A*, 105(29):9942–9947.

Biegert, A., Mayer, C., Remmert, M., Soeding, J., and Lupas, A. N. (2006). The mpi bioinformatics toolkit for protein sequence analysis. *Nucleic Acids Res*, 34(Web Server issue):W335–W339.

Biegert, A. and Soeding, J. (2009). Sequence context-specific profiles for homology searching. *Proc Natl Acad Sci U S A*, 106(10):3770–3775.

Bork, P., Holm, L., and Sander, C. (1994). The immunoglobulin fold. structural classification, sequence patterns and common core. *J Mol Biol*, 242(4):309–320.

Botos, I., Melnikov, E. E., Cherry, S., Khalatova, A. G., Rasulova, F. S., Tropea, J. E., Maurizi, M. R., Rotanova, T. V., Gustchina, A., and Wlodawer, A. (2004). Crystal structure of the aaa+ alpha domain of e. coli lon protease at 1.9a resolution. *J Struct Biol*, 146(1-2):113–122.

Branden, C. and Tooze, J. (1999). *Introduction to protein structure*. Garland Publishing.

Brennan, R. G. (1993). The winged-helix dna-binding motif: another helix-turn-helix takeoff. *Cell*, 74(5):773–776.

Brennan, R. G. and Matthews, B. W. (1989). The helix-turn-helix dna binding motif. *J Biol Chem*, 264(4):1903–1906.

Brenner, S. E., Koehl, P., and Levitt, M. (2000). The astral compendium for protein structure and sequence analysis. *Nucleic Acids Res*, 28(1):254–256.

Bull, A. T., Goodfellow, M., and Slater, J. H. (1992). Biodiversity as a source of innovation in biotechnology. *Annu Rev Microbiol*, 46:219–252.

Burley, S. K. and Bonanno, J. B. (2002). Structuring the universe of proteins. *Annu Rev Genomics Hum Genet*, 3:243–262.

Bystroff, C. and Baker, D. (1998). Prediction of local structure in proteins using a library of sequence-structure motifs. *J Mol Biol*, 281(3):565–577.

Caetano-Anolls, G., Kim, H. S., and Mittenthal, J. E. (2007). The origin of modern metabolic networks inferred from phylogenomic analysis of protein architecture. *Proc Natl Acad Sci U S A*, 104(22):9358–9363.

Caetano-Anolls, G., Wang, M., Caetano-Anolls, D., and Mittenthal, J. E. (2009). The origin, evolution and structure of the protein world. *Biochem J*, 417(3):621–637.

Castillo, R. M., Mizuguchi, K., Dhanaraj, V., Albert, A., Blundell, T. L., and Murzin, A. G. (1999). A six-stranded double-psi beta barrel is shared by several protein superfamilies. *Structure*, 7(2):227–236.

Challis, J. H. (1995). A procedure for determining rigid body transformation parameters. *J Biomech*, 28(6):733–737.

Chandonia, J.-M. and Brenner, S. E. (2006). The impact of structural genomics: expectations and outcomes. *Science*, 311(5759):347–351.

Chaudhuri, I., Soeding, J., and Lupas, A. N. (2008). Evolution of the beta-propeller fold. *Proteins*, 71(2):795–803.

Chen, J. C., Krucinski, J., Miercke, L. J., Finer-Moore, J. S., Tang, A. H., Leavitt, A. D., and Stroud, R. M. (2000). Crystal structure of the hiv-1 integrase catalytic core and c-terminal domains: a model for viral dna binding. *Proc Natl Acad Sci U S A*, 97(15):8233–8238.

Cheng, H., Kim, B.-H., and Grishin, N. V. (2008). Malisam: a database of structurally analogous motifs in proteins. *Nucleic Acids Res*, 36(Database issue):D211–D217.

Chou, K. C. and Zhang, C. T. (1995). Prediction of protein structural classes. *Crit Rev Biochem Mol Biol*, 30(4):275–349.

Coles, M., Diercks, T., Liermann, J., Groeger, A., Rockel, B., Baumeister, W., Koretke, K. K., Lupas, A., Peters, J., and Kessler, H. (1999). The solution structure of vat-n reveals a 'missing link' in the evolution of complex enzymes from a simple betaalphabetabeta element. *Curr Biol*, 9(20):1158–1168.

Coles, M., Djuranovic, S., Soeding, J., Frickey, T., Koretke, K., Truffault, V., Martin, J., and Lupas, A. N. (2005). Abrb-like transcription factors assume a swapped hairpin fold that is evolutionarily related to double-psi beta barrels. *Structure*, 13(6):919–928.

Coles, M., Hulko, M., Djuranovic, S., Truffault, V., Koretke, K., Martin, J., and Lupas, A. N. (2006). Common evolutionary origin of swapped-hairpin and double-psi beta barrels. *Structure*, 14(10):1489–1498.

Copley, R. R. and Bork, P. (2000). Homology among (betaalpha)(8) barrels: implications for the evolution of metabolic pathways. *J Mol Biol*, 303(4):627–641.

Copley, R. R., Russell, R. B., and Ponting, C. P. (2001). Sialidase-like asp-boxes: sequence-similar structures within different protein folds. *Protein Sci*, 10(2):285–292.

Cordes, M. H., Davidson, A. R., and Sauer, R. T. (1996). Sequence space, folding and protein design. *Curr Opin Struct Biol*, 6(1):3–10.

Cubonov, L., Sandman, K., Hallam, S. J., Delong, E. F., and Reeve, J. N. (2005). Histones in crenarchaea. *J Bacteriol*, 187(15):5482–5485.

Cuff, A., Redfern, O. C., Greene, L., Sillitoe, I., Lewis, T., Dibley, M., Reid, A., Pearl, F., Dallman, T., Todd, A., Garratt, R., Thornton, J., and Orengo, C. (2009). The cath hierarchy revisited-structural divergence in domain superfamilies and the continuity of fold space. *Structure*, 17(8):1051–1062.

Cuff, A. L., Sillitoe, I., Lewis, T., Clegg, A. B., Rentzsch, R., Furnham, N., Pellegrini-Calace, M., Jones, D., Thornton, J., and Orengo, C. A. (2011). Extending cath: increasing coverage of the protein structure universe and linking structure with function. *Nucleic Acids Res*, 39(Database issue):D420–D426.

D'Andrea, L. D. and Regan, L. (2003). Tpr proteins: the versatile helix. *Trends Biochem Sci*, 28(12):655–662.

Day, R., Beck, D. A. C., Armen, R. S., and Daggett, V. (2003). A consensus view of fold space: combining scop, cath, and the dali domain dictionary. *Protein Sci*, 12(10):2150–2160.

Diemand, A. V. and Lupas, A. N. (2006). Modeling aaa+ ring complexes from monomeric structures. *J Struct Biol*, 156(1):230–243.

Dietmann, S., Park, J., Notredame, C., Heger, A., Lappe, M., and Holm, L. (2001). A fully automatic evolutionary classification of protein folds: Dali domain dictionary version 3. *Nucleic Acids Res*, 29(1):55–57.

Doherty, A. J., Serpell, L. C., and Ponting, C. P. (1996). The helix-hairpin-helix dna-binding motif: a structural basis for non-sequence-specific recognition of dna. *Nucleic Acids Res*, 24(13):2488–2497.

Dominguez, C., Boelens, R., and Bonvin, A. M. J. J. (2003). Haddock: a protein-protein docking approach based on biochemical or biophysical information. *J Am Chem Soc*, 125(7):1731–1737.

Doudna, J. A. and Cech, T. R. (2002). The chemical repertoire of natural ribozymes. *Nature*, 418(6894):222–228.

Drlica, K. and Rouviere-Yaniv, J. (1987). Histonelike proteins of bacteria. *Microbiol Rev*, 51:301–319.

Eddy, S. R. (1998). Profile hidden markov models. *Bioinformatics*, 14(9):755–763.

Fahrner, R. L., Cascio, D., Lake, J. A., and Slesarev, A. (2001). An ancestral nuclear protein assembly: Crystal structure of the Methanopyrus kandleri histone. *Protein Sci*, 10:2002–2007.

Fetrow, J. S. and Godzik, A. (1998). Function driven protein evolution. a possible proto-protein for the rna-binding proteins. *Pac Symp Biocomput*, pages 485–496.

Finkelstein, A. V. and Ptitsyn, O. B. (1987). Why do globular proteins fit the limited set of folding patterns? *Prog Biophys Mol Biol*, 50(3):171–190.

Frickey, T. and Lupas, A. (2004). Clans: a java application for visualizing protein families based on pairwise similarity. *Bioinformatics*, 20(18):3702–3704.

Friedberg, I. and Godzik, A. (2005). Connecting the protein structure universe by using sparse recurring fragments. *Structure*, 13(8):1213–1224.

Gao, H., Cervantes, R. B., Mandell, E. K., Otero, J. H., and Lundblad, V. (2007). Rpa-like proteins mediate yeast telomere function. *Nat Struct Mol Biol*, 14(3):208–214.

Gay, N. J. and Walker, J. E. (1983). Homology between human bladder carcinoma oncogene product and mitochondrial atp-synthase. *Nature*, 301(5897):262–264.

GenBank (2012). `http://www.ncbi.nlm.nih.gov/genbank/`.

Glansdorff, N., Xu, Y., and Labedan, B. (2008). The last universal common ancestor: emergence, constitution and genetic legacy of an elusive forerunner. *Biol Direct*, 3:29.

Goebl, M. and Yanagida, M. (1991). The tpr snap helix: a novel protein repeat motif from mitosis to transcription. *Trends Biochem Sci*, 16(5):173–177.

GOLD (2012). `http://www.genomesonline.org/`.

Gotte, G., Bertoldi, M., and Libonati, M. (1999). Structural versatility of bovine ribonuclease a. distinct conformers of trimeric and tetrameric aggregates of the enzyme. *Eur J Biochem*, 265(2):680–687.

Green, S. M., Gittis, A. G., Meeker, A. K., and Lattman, E. E. (1995). One-step evolution of a dimer from a monomeric protein. *Nat Struct Biol*, 2(9):746–751.

Greene, L. H., Lewis, T. E., Addou, S., Cuff, A., Dallman, T., Dibley, M., Redfern, O., Pearl, F., Nambudiry, R., Reid, A., Sillitoe, I., Yeats, C., Thornton, J. M., and Orengo, C. A. (2007). The cath domain structure database: new protocols and classification levels give a more comprehensive resource for exploring evolution. *Nucleic Acids Res*, 35(Database issue):D291–D297.

Grishin, N. V. (2000). C-terminal domains of escherichia coli topoisomerase i belong to the zinc-ribbon superfamily. *J Mol Biol*, 299(5):1165–1177.

Grishin, N. V. (2001a). Fold change in evolution of protein structures. *J Struct Biol*, 134(2-3):167–185.

Grishin, N. V. (2001b). Kh domain: one motif, two folds. *Nucleic Acids Res*, 29(3):638–643.

Grishin, N. V. (2001c). Treble clef finger–a functionally diverse zinc-binding structural motif. *Nucleic Acids Res*, 29(8):1703–1714.

Guex, N. and Peitsch, M. C. (1997). Swiss-model and the swiss-pdbviewer: an environment for comparative protein modeling. *Electrophoresis*, 18(15):2714–2723.

Guo, F. and Adhya, S. (2007). Spiral structure of escherichia coli hualphabeta provides foundation for dna supercoiling. *Proc Natl Acad Sci U S A*, 104(11):4309–4314.

Guo, Z. and Eisenberg, D. (2006). Runaway domain swapping in amyloid-like fibrils of t7 endonuclease i. *Proc Natl Acad Sci U S A*, 103(21):8042–8047.

Hadjithomas, M. and Moudrianakis, E. N. (2011). Experimental evidence for the role of domain swapping in the evolution of the histone fold. *Proc Natl Acad Sci U S A*, 108:13462–13467.

Hoecker, B., Claren, J., and Sterner, R. (2004). Mimicking enzyme evolution by generating new (betaalpha)8-barrels from (betaalpha)4-half-barrels. *Proc Natl Acad Sci U S A*, 101(47):16448–16453.

Hoecker, B., Schmidt, S., and Sterner, R. (2002). A common evolutionary origin of two elementary enzyme folds. *FEBS Lett*, 510(3):133–135.

Holm, L. and Sander, C. (1993). Protein structure comparison by alignment of distance matrices. *J Mol Biol*, 233(1):123–138.

Hou, J., Jun, S.-R., Zhang, C., and Kim, S.-H. (2005). Global mapping of the protein structure space and application in structure-based inference of protein function. *Proc Natl Acad Sci U S A*, 102(10):3651–3656.

Hou, J., Sims, G. E., Zhang, C., and Kim, S.-H. (2003). A global representation of the protein fold space. *Proc Natl Acad Sci U S A*, 100(5):2386–2390.

Hulko, M., Lupas, A. N., and Martin, J. (2007). Inherent chaperone-like activity of aspartic proteases reveals a distant evolutionary relation to double-psi barrel domains of aaa-atpases. *Protein Sci*, 16(4):644–653.

Hunter, S., Apweiler, R., Attwood, T. K., Bairoch, A., Bateman, A., Binns, D., Bork, P., Das, U., Daugherty, L., Duquenne, L., Finn, R. D., Gough, J., Haft, D., Hulo, N., Kahn, D., Kelly, E., Laugraud, A., Letunic, I., Lonsdale, D., Lopez, R., Madera, M., Maslen, J., McAnulla, C., McDowall, J., Mistry, J., Mitchell, A., Mulder, N., Natale, D., Orengo, C., Quinn, A. F., Selengut, J. D., Sigrist, C. J. A., Thimma, M., Thomas, P. D., Valentin, F., Wilson, D., Wu, C. H., and Yeats, C. (2009). Interpro: the integrative protein signature database. *Nucleic Acids Res*, 37(Database issue):D211–D215.

Hutchinson, E. G. and Thornton, J. M. (1994). A revised set of potentials for beta-turn formation in proteins. *Protein Sci*, 3(12):2207–2216.

Hutchinson, E. G. and Thornton, J. M. (1996). Promotif–a program to identify and analyze structural motifs in proteins. *Protein Sci*, 5(2):212–220.

Izzo, A., Kamieniarz, K., and Schneider, R. (2008). The histone h1 family: specific members, specific functions? *Biol Chem*, 389(4):333–343.

Janowski, R., Kozak, M., Jankowska, E., Grzonka, Z., Grubb, A., Abrahamson, M., and Jaskolski, M. (2001). Human cystatin c, an amyloidogenic protein, dimerizes through three-dimensional domain swapping. *Nat Struct Biol*, 8(4):316–320.

Jeffares, D. C., Poole, A. M., and Penny, D. (1998). Relics from the rna world. *J Mol Evol*, 46(1):18–36.

Jones, D. T. (1999). Protein secondary structure prediction based on position-specific scoring matrices. *J Mol Biol*, 292(2):195–202.

Joyce, G. F. (2002). The antiquity of rna-based evolution. *Nature*, 418(6894):214–221.

Kabsch, W. and Sander, C. (1983). Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, 22(12):2577–2637.

Kano, Y., Ogawa, T., Ogura, T., Hiraga, S., Okazaki, T., and Imamoto, F. (1991). Participation of the histone-like protein hu and of ihf in minichromosomal maintenance in escherichia coli. *Gene*, 103(1):25–30.

Keefe, A. D. and Szostak, J. W. (2001). Functional proteins from a random-sequence library. *Nature*, 410(6829):715–718.

Keller, M., Bloechl, E., Waechtershaeuser, G., and Stetter, K. O. (1994). Formation of amide bonds without a condensation agent and implications for origin of life. *Nature*, 368(6474):836–838.

Kinch, L. N. and Grishin, N. V. (2002). Evolution of protein structures and functions. *Curr Opin Struct Biol*, 12(3):400–408.

Kolodny, R., Petrey, D., and Honig, B. (2006). Protein structure comparison: implications for the nature of 'fold space', and structure and function prediction. *Curr Opin Struct Biol*, 16(3):393–398.

Kopec, K. O., Alva, V., and Lupas, A. N. (2010). Homology of smp domains to the tulip superfamily of lipid-binding proteins provides a structural basis for lipid exchange between er and mitochondria. *Bioinformatics*, 26(16):1927–1931.

Kornberg, R. D. and Thomas, J. O. (1974). Chromatin structure; oligomers of the histones. *Science*, 184:865–868.

Kortt, A. A., Malby, R. L., Caldwell, J. B., Gruen, L. C., Ivancic, N., Lawrence, M. C., Howlett, G. J., Webster, R. G., Hudson, P. J., and Colman, P. M. (1994). Recombinant anti-sialidase single-chain variable fragment antibody. characterization, formation of dimer and higher-molecular-mass multimers and the solution of the crystal structure of the single-chain variable fragment/sialidase complex. *Eur J Biochem*, 221(1):151–157.

Kryshtafovych, A., Venclovas, C., Fidelis, K., and Moult, J. (2005). Progress over the first decade of casp experiments. *Proteins*, 61 Suppl 7:225–236.

Lee, A. Y.-L., Hsu, C.-H., and Wu, S.-H. (2004). Functional domains of brevibacillus thermoruber lon protease for oligomerization and dna binding: role of n-terminal and sensor and substrate discrimination domains. *J Biol Chem*, 279(33):34903–34912.

Lee, S. and Tsai, F. T. F. (2005). Molecular chaperones in protein quality control. *J Biochem Mol Biol*, 38(3):259–265.

Levinthal, C. (1968). Are there pathways for protein folding? *J Chim Phys*, 65:44–45.

Li, W. T., Sandman, K., Pereira, S. L., and Reeve, J. N. (2000). Mj1647, an open reading frame in the genome of the hyperthermophile methanococcus jannaschii, encodes a very thermostable archaeal histone with a c-terminal extension. *Extremophiles*, 4(1):43–51.

Liang, P.-H., Ko, T.-P., and Wang, A. H.-J. (2002). Structure, mechanism and function of prenyltransferases. *Eur J Biochem*, 269(14):3339–3354.

Liu, Y. and Eisenberg, D. (2002). 3d domain swapping: as domains continue to swap. *Protein Sci*, 11(6):1285–1299.

Luger, K., Mäder, A. W., Richmond, R. K., Sargent, D. F., and Richmond, T. J. (1997). Crystal structure of the nucleosome core particle at 2.8 A resolution. *Nature*, 389:251–260.

Luger, K. and Richmond, T. J. (1998). The histone tails of the nucleosome. *Curr Opin Genet Dev*, 8(2):140–146.

Lukatsky, D. B., Shakhnovich, B. E., Mintseris, J., and Shakhnovich, E. I. (2007). Structural similarity enhances interaction propensity of proteins. *J Mol Biol*, 365(5):1596–1606.

Lupas, A. N. (2008). The long coming of computational structural biology. *J Struct Biol*, 163(3):254–257.

Lupas, A. N., Ponting, C. P., and Russell, R. B. (2001). 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*, 134:191–203.

Lynch, T. W., Read, E. K., Mattis, A. N., Gardner, J. F., and Rice, P. A. (2003). Integration host factor: putting a twist on protein-dna recognition. *J Mol Biol*, 330(3):493–502.

Madej, T., Addess, K. J., Fong, J. H., Geer, L. Y., Geer, R. C., Lanczycki, C. J., Liu, C., Lu, S., Marchler-Bauer, A., Panchenko, A. R., Chen, J., Thiessen, P. A., Wang, Y., Zhang, D., and Bryant, S. H. (2012).

Mmdb: 3d structures and macromolecular interactions. *Nucleic Acids Res*, 40(Database issue):D461–D464.

Marsden, R. L., Lewis, T. A., and Orengo, C. A. (2007). Towards a comprehensive structural coverage of completed genomes: a structural genomics viewpoint. *BMC Bioinformatics*, 8:86.

Matte, A., Goldie, H., Sweet, R. M., and Delbaere, L. T. (1996). Crystal structure of escherichia coli phosphoenolpyruvate carboxykinase: a new structural family with the p-loop nucleoside triphosphate hydrolase fold. *J Mol Biol*, 256(1):126–143.

Maurizi, M. R. and Xia, D. (2004). Protein binding and disruption by clp/hsp100 chaperones. *Structure*, 12(2):175–183.

McDonald, I. K. and Thornton, J. M. (1994). Satisfying hydrogen bonding potential in proteins. *J Mol Biol*, 238(5):777–793.

McLachlan, A. D. (1979). Three-fold structural pattern in the soybean trypsin inhibitor (kunitz). *J Mol Biol*, 133(4):557–563.

McLachlan, A. D. (1980). Repeated folding pattern in copper-zinc superoxide dismutase. *Nature*, 285(5762):267–268.

McLachlan, A. D., Bloomer, A. C., and Butler, P. J. (1980). Structural repeats and evolution of tobacco mosaic virus coat protein and rna. *J Mol Biol*, 136(3):203–224.

Minor, D. L. and Kim, P. S. (1996). Context-dependent secondary structure formation of a designed protein sequence. *Nature*, 380(6576):730–734.

Murray, A. J., Lewis, S. J., Barclay, A. N., and Brady, R. L. (1995). One sequence, two folds: a metastable structure of cd2. *Proc Natl Acad Sci U S A*, 92(16):7337–7341.

Mkiniemi, M., Pospiech, H., Kilpelinen, S., Jokela, M., Vihinen, M., and Syvoja, J. E. (1999). A novel family of dna-polymerase-associated b subunits. *Trends Biochem Sci*, 24(1):14–16.

Nagano, N., Orengo, C. A., and Thornton, J. M. (2002). One fold with many functions: the evolutionary relationships between tim barrel families based on their sequences, structures and functions. *J Mol Biol*, 321(5):741–765.

Neuwald, A. F., Aravind, L., Spouge, J. L., and Koonin, E. V. (1999). Aaa+: A class of chaperone-like atpases associated with the assembly, operation, and disassembly of protein complexes. *Genome Res*, 9(1):27–43.

Nilges, M. (1995). Calculation of protein structures with ambiguous distance restraints. automated assignment of ambiguous noe crosspeaks and disulphide connectivities. *J Mol Biol*, 245(5):645–660.

Nuutinen, T., Tossavainen, H., Fredriksson, K., Piril, P., Permi, P., Pospiech, H., and Syvaoja, J. E. (2008). The solution structure of the amino-terminal domain of human dna polymerase epsilon subunit b is homologous to c-domains of aaa+ proteins. *Nucleic Acids Res*, 36(15):5102–5110.

Ogihara, N. L., Ghirlanda, G., Bryson, J. W., Gingery, M., DeGrado, W. F., and Eisenberg, D. (2001). Design of three-dimensional domain-swapped dimers and fibrous oligomers. *Proc Natl Acad Sci U S A*, 98(4):1404–1409.

Ogura, T., Whiteheart, S. W., and Wilkinson, A. J. (2004). Conserved arginine residues implicated in atp hydrolysis, nucleotide-sensing, and inter-subunit interactions in aaa and aaa+ atpases. *J Struct Biol*, 146(1-2):106–112.

Orengo, C. A., Flores, T. P., Taylor, W. R., and Thornton, J. M. (1993). Identification and classification of protein fold families. *Protein Eng*, 6(5):485–500.

Orengo, C. A., Jones, D. T., and Thornton, J. M. (1994). Protein superfamilies and domain superfolds. *Nature*, 372(6507):631–634.

Orengo, C. A. and Thornton, J. M. (2005). Protein families and their evolution-a structural perspective. *Annu Rev Biochem*, 74:867–900.

Otaka, E. and Ooi, T. (1987). Examination of protein sequence homologies: Iv. twenty-seven bacterial ferredoxins. *J Mol Evol*, 26(3):257–267.

Pace, C. N., Shirley, B. A., McNutt, M., and Gajiwala, K. (1996). Forces contributing to the conformational stability of proteins. *FASEB J*, 10(1):75–83.

Pascual-Garca, A., Abia, D., Ortiz, A. R., and Bastolla, U. (2009). Cross-over between discrete and continuous protein structure space: insights into automatic classification and networks of protein structures. *PLoS Comput Biol*, 5(3):e1000331.

Pauling, L., Corey, R. B., and Branson, H. R. (1951). The structure of proteins; two hydrogen-bonded helical configurations of the polypeptide chain. *Proc Natl Acad Sci U S A*, 37(4):205–211.

PDB (2012). `http://www.pdb.org/`.

Pearson, W. R. and Lipman, D. J. (1988). Improved tools for biological sequence comparison. *Proc Natl Acad Sci U S A*, 85(8):2444–2448.

Pereira, S. L., Grayling, R. A., Lurz, R., and Reeve, J. N. (1997). Archaeal nucleosomes. *Proc Natl Acad Sci U S A*, 94:12633–12637.

Perisic, O., Webb, P. A., Holliger, P., Winter, G., and Williams, R. L. (1994). Crystal structure of a diabody, a bivalent antibody fragment. *Structure*, 2(12):1217–1226.

Perutz, M. F., Rossmann, M. G., Cullis, A. F., Muirhead, H., Will, G., and North, A. C. (1960). Structure of haemoglobin: a three-dimensional fourier synthesis at 5.5-a. resolution, obtained by x-ray analysis. *Nature*, 185(4711):416–422.

Piccoli, R., Donato, A. D., and D'Alessio, G. (1988). Co-operativity in seminal ribonuclease function. kinetic studies. *Biochem J*, 253(2):329–336.

Ponting, C. P. and Russell, R. B. (2000). Identification of distant homologues of fibroblast growth factors suggests a common ancestor for all beta-trefoil proteins. *J Mol Biol*, 302(5):1041–1047.

Ponting, C. P. and Russell, R. R. (2002). The natural history of protein domains. *Annu Rev Biophys Biomol Struct*, 31:45–71.

Punta, M., Coggill, P. C., Eberhardt, R. Y., Mistry, J., Tate, J., Boursnell, C., Pang, N., Forslund, K., Ceric, G., Clements, J., Heger, A., Holm, L., Sonnhammer, E. L. L., Eddy, S. R., Bateman, A., and Finn, R. D. (2012). The pfam protein families database. *Nucleic Acids Res*, 40(D1):D290–D301.

Qiu, Y., Tereshko, V., Kim, Y., Zhang, R., Collart, F., Yousef, M., Kossiakoff, A., and Joachimiak, A. (2006). The crystal structure of Aq_328 from the hyperthermophilic bacteria Aquifex aeolicus shows an ancestral histone fold. *Proteins*, 62:8–16.

Ramakrishnan, V., Finch, J. T., Graziano, V., Lee, P. L., and Sweet, R. M. (1993). Crystal structure of globular domain of histone h5 and its implications for nucleosome binding. *Nature*, 362(6417):219–223.

Rao, S. T. and Rossmann, M. G. (1973). Comparison of super-secondary structures in proteins. *J Mol Biol*, 76(2):241–256.

RefSeq (2012). `http://www.ncbi.nlm.nih.gov/RefSeq/`.

Remmert, M., Linke, D., Lupas, A. N., and Soeding, J. (2009). Hhomp–prediction and classification of outer membrane proteins. *Nucleic Acids Res*, 37(Web Server issue):W446–W451.

Reynaud, E. (2010). Protein misfolding and degenerative diseases. *Nature Education 3(9):28.*

Rost, B. (2002). Did evolution leap to create the protein universe? *Curr Opin Struct Biol*, 12(3):409–416.

Rouvire-Yaniv, J., Yaniv, M., and Germond, J. E. (1979). E. coli dna binding protein hu forms nucleosomelike structure with circular double-stranded dna. *Cell*, 17(2):265–274.

Sadreyev, R. and Grishin, N. (2003). Compass: a tool for comparison of multiple protein alignments with assessment of statistical significance. *J Mol Biol*, 326(1):317–336.

Sadreyev, R. I., Kim, B.-H., and Grishin, N. V. (2009). Discrete-continuous duality of protein structure space. *Curr Opin Struct Biol*, 19(3):321–328.

Salem, G. M., Hutchinson, E. G., Orengo, C. A., and Thornton, J. M. (1999). Correlation of observed fold frequency with the occurrence of local structural motifs. *J Mol Biol*, 287(5):969–981.

Sambashivan, S., Liu, Y., Sawaya, M. R., Gingery, M., and Eisenberg, D. (2005). Amyloid-like fibrils of ribonuclease a with three-dimensional domain-swapped and native-like structure. *Nature*, 437(7056):266–269.

Sandman, K., Krzycki, J. A., Dobrinski, B., Lurz, R., and Reeve, J. N. (1990). Hmf, a dna-binding protein isolated from the hyperthermophilic archaeon methanothermus fervidus, is most closely related to histones. *Proc Natl Acad Sci U S A*, 87(15):5788–5791.

Sandman, K. and Reeve, J. N. (2001). Chromosome packaging by archaeal histones. *Adv Appl Microbiol*, 50:75–99.

Sandman, K. and Reeve, J. N. (2006). Archaeal histones and the origin of the histone fold. *Curr Opin Microbiol*, 9(5):520–525.

Schymkowitz, J. W., Rousseau, F., Wilkinson, H. R., Friedler, A., and Itzhaki, L. S. (2001). Observation of signal transduction in three-dimensional domain swapping. *Nat Struct Biol*, 8(10):888–892.

Simons, K. T., Bonneau, R., Ruczinski, I., and Baker, D. (1999). Ab initio protein structure prediction of casp iii targets using rosetta. *Proteins*, Suppl 3:171–176.

Soeding, J. (2005). Protein homology detection by hmm-hmm comparison. *Bioinformatics*, 21(7):951–960.

Soeding, J. and Lupas, A. N. (2003). More than the sum of their parts: on the evolution of proteins from peptides. *Bioessays*, 25(9):837–846.

Soeding, J., Remmert, M., and Biegert, A. (2006). Hhrep: de novo protein repeat detection and the origin of tim barrels. *Nucleic Acids Res*, 34(Web Server issue):W137–W142.

Staker, B. L., Korber, P., Bardwell, J. C., and Saper, M. A. (2000). Structure of hsp15 reveals a novel rna-binding motif. *EMBO J*, 19(4):749–757.

Taylor, W. R. (2007). Evolutionary transitions in protein fold space. *Curr Opin Struct Biol*, 17(3):354–361.

Tian, B., Bevilacqua, P. C., Diegelman-Parente, A., and Mathews, M. B. (2004). The double-stranded-rna-binding motif: interference and much more. *Nat Rev Mol Cell Biol*, 5(12):1013–1023.

Voet, D. and Voet, J. G. (2004). *Biochemistry*, volume 1. Wiley: Hoboken, NJ., 3 edition.

Wheelan, S. J., Marchler-Bauer, A., and Bryant, S. H. (2000). Domain size distributions can predict domain boundaries. *Bioinformatics*, 16(7):613–618.

White, M. F. and Bell, S. D. (2002). Holding it together: chromatin in the archaea. *Trends Genet*, 18(12):621–626.

Wieden, H. J., Wintermeyer, W., and Rodnina, M. V. (2001). A common structural motif in elongation factor ts and ribosomal protein l7/12 may be involved in the interaction with elongation factor tu. *J Mol Evol*, 52(2):129–136.

Wierenga, R. K. (2001). The tim-barrel fold: a versatile framework for efficient enzymes. *FEBS Lett*, 492(3):193–198.

Worbs, M., Bourenkov, G. P., Bartunik, H. D., Huber, R., and Wahl, M. C. (2001). An extended rna binding surface through arrayed s1 and kh domains in transcription factor nusa. *Mol Cell*, 7(6):1177–1189.

Xie, L. and Bourne, P. E. (2008). Detecting evolutionary relationships across existing fold space, using sequence order-independent profile-profile alignments. *Proc Natl Acad Sci U S A*, 105(14):5441–5446.

Zarbock, J., Clore, G. M., and Gronenborn, A. M. (1986). Nuclear magnetic resonance study of the globular domain of chicken histone h5: resonance assignment and secondary structure. *Proc Natl Acad Sci U S A*, 83(20):7628–7632.

Zeth, K., Ravelli, R. B., Paal, K., Cusack, S., Bukau, B., and Dougan, D. A. (2002). Structural analysis of the adaptor protein clps in complex with the n-terminal domain of clpa. *Nat Struct Biol*, 9(12):906–911.

Zhang, J. G., Farley, A., Nicholson, S. E., Willson, T. A., Zugaro, L. M., Simpson, R. J., Moritz, R. L., Cary, D., Richardson, R., Hausmann, G., Kile, B. J., Kent, S. B., Alexander, W. S., Metcalf, D., Hilton, D. J., Nicola, N. A., and Baca, M. (1999). The conserved socs box motif in suppressors of cytokine signaling binds to elongins b and c and may couple bound proteins to proteasomal degradation. *Proc Natl Acad Sci U S A*, 96(5):2071–2076.

Zhang, Y. and Skolnick, J. (2005). Tm-align: a protein structure alignment algorithm based on the tm-score. *Nucleic Acids Res*, 33:2302–2309.