Structural and functional constraints in the evolution of protein families

Catherine L. Worth‡, Sungsam Gong* and Tom L. Blundell**

Abstract | High-throughput genomic sequencing has focused attention on understanding differences between species and between individuals. When this genetic variation affects protein sequences, the rate of amino acid substitution reflects both Darwinian selection for functionally advantageous mutations and selectively neutral evolution operating within the constraints of structure and function. During neutral evolution, whereby mutations accumulate by random drift, amino acid substitutions are constrained by factors such as the formation of intramolecular and intermolecular interactions and the accessibility to water or lipids surrounding the protein. These constraints arise from the need to conserve a specific architecture and to retain interactions that mediate functions in protein families and superfamilies.

Chaperone
A protein that assists in the folding or unfolding and the assembly or disassembly of other macromolecular structures.

Neutral drift
The process whereby random sampling effects over successive generations give rise to stochastic changes in the allele frequencies within a population.

β-lactamase
An enzyme produced by some bacteria that confers resistance to β-lactam antibiotics.

**Biochemistry Department, University of Cambridge, Cambridge, CB2 1GA, UK.*
‡Leibniz-Institut für Molekulare Pharmakologie, Campus Berlin-Buch, Berlin, 13125, Germany.
Correspondence to T.L.B.
e-mail:
tom@cryst.bioc.cam.ac.uk
doi:10.1038/nrm2762
Published online
16 September 2009

Although amino acid sequence determines three-dimensional protein structure — sometimes with a little help from chaperones — tertiary structure tends to be better conserved than sequence in evolution[1,2]. Thus, in homologous families of proteins, functions are often retained and structures are usually very similar even though sequences have diverged. This is even more evident in protein superfamilies, in which overall sequence similarity can be insignificant but structural and functional similarities still provide evidence of distant common ancestry.

Around 40 years ago Kimura and Ohta developed the neutral theory of evolution, which states that most evolutionary changes at the molecular level are caused by neutral drift — the acceptance of selectively neutral mutations[3,4]. They suggested that mutations that disrupt the existing structure and function of a molecule occur less frequently in evolution than neutral mutations. This was elaborated by Zuckerkandl and colleagues in the functional density hypothesis, which proposes that the rate of evolution is determined by the proportion of all the possible mutations that produce a protein that is functionally equivalent to the wild type[5,6]. More recently it was found that proteins with many interaction partners evolve more slowly than those with few interaction partners[7–9], but this has been disputed[9]. Analyses of the arrangements of polypeptide chains, often called protein folds, indicate that those that occur frequently tend to adopt regular architectures[10].

Insights into the effects of missense mutations on protein folding, structure and function, and thereby into the roles of wild-type amino acids, have been obtained from careful experimental approaches to amino acid substitution, such as site-specific mutagenesis. Such approaches dissect the various contributions of individual side chains in a systematic way. For example, the complex relationships between amino acid substitutions and the folding, stability and activity of proteins such as p53 have been explored by combining molecular biology and physical-organic chemistry[11–15]. The bacteriophage T4 lysozyme has been used as a model system to investigate the tolerance of proteins to amino acid replacement, insertion and deletion of both single amino acids and longer segments of the polypeptide chain using high resolution X-ray crystallography[16–19]. These classical studies have shown that a protein can tolerate substantial changes, consistent with the observations of evolving proteins. Similar experiments have investigated how mutations are tolerated in the active sites of enzymes; for example, the study of mutant β-lactamase enzymes in order to understand the mutant bacteria's resistance to penicillin analogues showed that as penicillins become larger the enzymes evolve larger active sites and become less stable[16]. This understanding has been exploited in the design of new inhibitors. The combination of molecular biology, organic chemistry and state-of-the-art high-throughput screening technologies in directed evolution to generate new proteins with tailor-made properties demonstrates that neutral drift can lead to more promiscuous enzymes with broader functions[17,18].

These experimental studies have provided invaluable quantitative information that is complementary to and largely consistent with the results of comparisons of the sequences and structures of protein families and

Box 1 | **Various constraints of protein evolution**

In this Review, we focus on local structural environments of amino acids as major constraints on the possible substitutions of amino acids during protein evolution. We also address the question of the importance of maintaining the function of a protein in imposing constraints, especially where molecular recognition is crucial, such as in enzyme active sites. However, there are many other constraints that are less well understood but provide important pressures in evolution. They include those that arise from DNA packaging and gene splicing and from the requirement for reliable and well-coordinated gene expression[94,95,97]. For example, ubiquitously expressed proteins tend to evolve slower than tissue-specific proteins. In addition, constraints arise from the process of protein folding[98,99], from the importance of retaining various conformational changes and flexibility that mediate functions in the cell and from the need to avoid opportunistic interactions (interactions occurring by chance) and amyloid formation — aggregation of misfolded proteins into a highly ordered fibril-like structure[100,101]. Furthermore, in order to prevent accumulation of damaging proteins the protein degradation system must be finely controlled, especially for misfolded proteins resulting from mutations[102]. Recently, it has been found that epigenetic factors, such as DNA methylation and chromatin remodelling, have important roles in the regulation of gene expression[103] that eventually affect the evolution of proteins. Hence, an integrated approach is required to comprehensively understand protein evolution[23].

superfamilies, which we now review. Such comparative analyses of proteins can throw light on these observations by focusing on substitutions at topologically equivalent amino acid positions in families and superfamilies and by integrating the information into local environment-dependent substitution tables. These show that identical amino acids are substituted in different ways, depending on the role of the amino acid in maintaining the protein's structure and functional interactions. What then is the nature of the constraints on amino acid substitutions that give rise to distinct patterns of protein evolution?

In this Review we consider amino acid substitutions that have occurred in protein families and superfamilies. We do not discuss the origins of folds or their evolution by additions and subtractions of elements of secondary structure, gene duplications and fusions; these have been widely reviewed elsewhere[19–22]. Neither do we consider constraints arising from the genomic position of the encoding genes, expression patterns, position in biological networks or robustness to translation[23] (see BOX 1 for various constraints of protein evolution). Rather, we focus on how the amino acid substitutions during divergent evolution of protein families are constrained by the structure and functional interactions of a protein.

We show that amino acid substitutions can be understood better and predicted more accurately if the three-dimensional environment of the amino acid side chain — known as the local structural environment — is defined in the functional state of the protein, for example in terms of secondary structure, accessibility to the water, lipids or other medium surrounding the protein and formation of hydrogen bonds. In particular, we focus on water-inaccessible polar side chains, which provide strong structural and functional constraints in the evolution of protein families. We show that these can give rise to characteristic architectural motifs resulting from their need to satisfy hydrogen bonding requirements.

Constraint
A structural and dynamic system, or functional factor, that influences the acceptance of amino acid substitutions that occur in divergent protein families. Given that selection occurs at the level of the organism and that individual proteins and the systems in which they evolve are plastic, these constraints tend not to 'force' but rather to 'restrain' the substitutions that occur in evolution.

Orthologues
Genes (or gene products) descended from a common ancestral origin that diverged as a result of a speciation event.

Hydrogen bonding potential
The capacity of atoms to act as proton donors or acceptors in the formation of hydrogen bonds.

Jelly roll
An eight-stranded β-sandwich that is formed by four Greek key motifs, each consisting of four sequential antiparallel β-strands.

β-propeller
An all-β protein architecture comprising four to eight blade-shaped β-sheets arranged toroidally around a central axis.

Comparative analyses of homologous proteins

We first try to understand family resemblances before we seek to recognize the unique features of individual family members. This is best achieved by comparing the sequences and structures of members of families and superfamilies — proteins that are homologous or descended from a common ancestor — to be found among the more than fifty thousand proteins for which architectures have been determined at high resolution. We can then define each amino acid position in a protein family in terms of its local structural environment and investigate how structural constraints affect the amino acid substitutions that have been accepted during evolution. One major challenge here is to distinguish orthologues, which have the same functions in different organisms, from paralogues, which result from gene duplication and might have evolved new functions[24]. For paralogues the constraints will have changed. Generally orthologues are defined on the basis of sequence similarity but this remains a source of uncertainty in comparative analyses.

The first comparisons 40 to 50 years ago of primary and tertiary structures of homologous proteins (globins, serine proteinases and lysozymes) focused on accessibility to water, usually called solvent accessibility, and showed that the solvent-inaccessible cores of proteins tended to be closely packed, more hydrophobic and more conserved than the surface regions[25]. Analyses of the structures from many protein families (BOX 2) show that this remains a useful generalization. These early analyses also focused on regular secondary structures, such as α-helices and β-sheets, which were immediately recognized to favour particular amino acids, so providing further constraints on evolutionary change[26–28].

Pauling and colleagues realized that the requirement for the satisfaction of the hydrogen bonding potential of polypeptide main-chain peptide amide (NH) and carbonyl (CO) groups would not only give rise to regular secondary structures[29,30] but also make the main chains of proteins more hydrophobic so that they could be buried in the core of a globular protein along with non-polar side chains. It soon became evident that these features of main-chain hydrogen bonding restrict protein architectures to a limited set of super-secondary structures formed by combining secondary structures into globular units, such as β-sandwiches and barrels, jelly rolls, β-propellers, α-helical bundles, αβ-Rossman folds, αβ-barrels and many others. Main-chain hydrogen bonding also has important roles in the formation of complex arches and turns that link α-helices and β-strands[31–33].

Nevertheless, many main-chain peptide CO and NH groups are left unsatisfied in their potential to form hydrogen bonds. An early analysis of hydrogen bonding revealed that ~40% of such groups do not form hydrogen bonds with main-chain atoms of other amino acids[34]. In general this lack of hydrogen bonding occurs at places where β-strands and α-helices terminate[34–38], bulge[39,40] or bend[41,42], but it is also common in polyproline or irregular, twisted β-strands[43,44] and in arches and turns[31–33,45,46]. The hydrogen bonding potential of these

Box 2 | **A selection of protein classification databases and similarity search servers**

Insight into evolutionary relationships can be gained by grouping similar proteins. Several classification resources categorize proteins based on their degree of similarity but they differ in definition and method. Nevertheless, there is general agreement on the hierarchical order of overall topology or fold, superfamily, family and individual domains. Many proteins with the same topology will have convergently evolved, but members of superfamilies and families are likely to have arisen from a common ancestor by divergent evolution. SCOP[104] and CATH[105] are two well-known databases of hierarchical protein structure classification. HOMSTRAD[70], PASS2 (REF. 106), Toccata[107] and Dali[108] provide superimposed and aligned protein families with various annotations at the residue level. CE[109] also provides structure comparison and alignment. MMDB provides structure–neighbour calculations such that each structure is linked to related three-dimensional domains[110]. Sequence-based protein family databases include Pfam[111] and InterPro[112]. InterPro is a consortium of several member databases such as PROSITE[113], Pfam, Prints[114], ProDom[115], SMART[116] and TIGRFAMs[117]. Using curated or computed protein classification schemes, homology detection can be achieved using sequence and/or structure similarity as implemented by Gene3D[118], Superfamily[119], PhyloFacts[120], CDD[121], PairsDB[122] and SMART. These databases and servers can be useful resources in the study of protein evolution and a comprehensive comparison of them is available in REF. 123.

motifs is satisfied by water molecules or by polar side chains; when the side chains are inaccessible they place a strong constraint on neutral drift.

Comparisons of homologous proteins show that interaction sites that mediate important functions by binding regulatory proteins, nucleic acids and other ligands also place strong evolutionary constraints on amino acid substitutions[47–50]. These interaction sites cannot be understood at the level of an isolated protein; rather, different proteins and sometimes other macromolecules associate to form a multicomponent system that serves as a functional unit and places significant constraints on evolutionary change. In insulin, for example, comparative analyses of family members have revealed that amino acid substitutions at the interfaces involved in dimer, hexamer and receptor complex formation have been under strong constraints since the evolution of bony fishes — only the rodent sub-order of Hystricomorpha, which includes animals such as the guinea pig and the coypu, has monomeric insulins[47]. Although the amino acid substitutions that lead to the loss of the ability of insulin to hexamerize in Hystricomorpha were first thought to be selectively neutral, it is now thought that they were probably selectively advantageous and provided a means of stably storing insulin, possibly in an environment with a shortage of zinc that prevented the use of zinc insulin hexamers as found in other mammals.

For enzymes, it is clear that the local environment of catalytic residues in reaction intermediates and transition states must be considered. The need for particular recognition sequences at sites of post-translational modification, of adaptor–template protein interactions and of allosteric effector binding also provides strong constraints. Recently, it has also become evident that many of these sites of molecular interaction or recognition lead to further constraints on the substitution of amino acid residues in the vicinity of protein binding sites but not in the immediate contact with a ligand[51].

Conservation and local environment

Sequence alignments of homologues of known structure can be used to help quantify the constraints that arise from both protein structure and function in a family of proteins. By defining the local structural environment of amino acid residues (secondary structure, solvent accessibility and formation of hydrogen bonds), distinct patterns of substitutions have been observed[52,53]. Environment-specific substitution tables (ESSTs) store these substitution data quantitatively in the form of probabilities and thereby provide information on the existence of each amino acid in a particular environment and the probability of it being substituted by any other amino acid (BOX 3).

These ESSTs show that amino acids with side chains that are hydrogen bonded to main-chain NH and CO groups are more conserved than those with side chains that are hydrogen bonded to other side chains. This is particularly evident when side chains are inaccessible to the solvent and when they form hydrogen bonds to main-chain NH groups. This implies that a crucial element in protein structure is the satisfaction of the hydrogen bond donor and acceptor properties of the main-chain NH and CO groups when the protein is folded. When these requirements are not satisfied by secondary structures, hydrogen bonds to side chains might be conserved to meet this requirement.

Solvent accessibility has a major role. It has long been known that residue conservation in the solvent-inaccessible regions is much higher than in those regions that are solvent accessible[54]. FIGURE 1 shows the clustering of 64 local structural environments with the UPGMA (unweighted pair group method with arithmetic mean) algorithm[55], based on distances among 64 substitution tables (64 × 64 distance matrix), to identify the structural constraints that determine the substitution patterns of amino acids. The distance between two substitution tables was measured by summing the differences in the probability of amino acid substitutions. The matrices for the 64 environments form 3 distinct clusters: 2 are distinguished by solvent accessibility (clusters 1 and 2 in FIG. 1), whereas the third is characterized by the presence of a positive φ main-chain torsion angle (cluster 3 in FIG. 1).

Even in the cluster of environments with positive φ main-chain torsion angles (see below), solvent accessibility divides the environments into two: accessible and inaccessible. Solvent inaccessibility thus puts constraints on the acceptance of selectively neutral

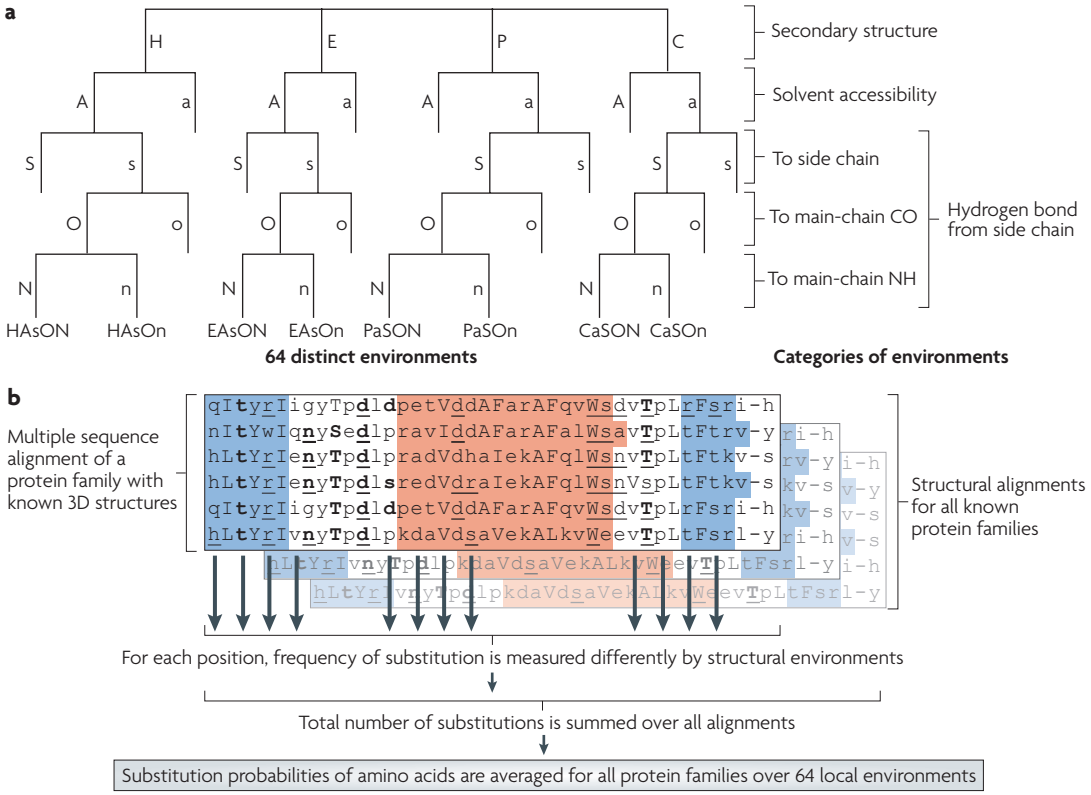
α-helical bundle
A protein fold consisting of multiple α-helices that are approximately parallel to one another.

αβ-Rossman fold
Two repeating β–α–β super-secondary motifs.

Distance matrix
An $n \times n$ array that represents the distances between a set of n elements.

Positive φ main-chain torsion angle
A positive dihedral angle around the nitrogen–α-carbon bonds in the protein main chain. For L-amino acids these bond angles are generally restricted to a negative value owing to steric hindrance from the side chains, but they can be positive when there is no side chain (Gly) or when polar side-chain interactions with the main-chain peptide units stabilize this conformation.

Box 3 | **Local structural environments and calculation of substitution tables**



Environment-specific substitution tables (ESSTs) provide the basic evidence that amino acid substitutions are constrained in different ways in different local environments. Such tables exploit categories of amino acid local structural environments, such as main-chain conformation and secondary structure, solvent accessibility, and hydrogen bonding between side chains and either main-chain groups or other side chains.

For example, in part **a** of the figure, amino acids can be classified into 1 of 64 environments: 4 from secondary structure (α-helix (H), β-strand (E), positive φ main-chain torsion angle (P) and coil (C)), 2 from solvent accessibility (accessible (A) and inaccessible (a)), and 8 from the existence (upper case) or absence (lower case) of hydrogen bonds from a side chain to another side chain (S and s), to a main-chain carbonyl group (O and o) and to a main-chain amide group (N and n). These combinations of structural features influence the substitution of amino acids and give rise to distinct patterns of amino acid substitutions. Part **b** of the figure shows how ESSTs can be generated from homologous protein structure alignments in which each residue has been annotated with three-dimensional (3D) structural features (explained above) and assigned to one of the 64 environments in JOY[124] format: solvent inaccessible (upper case), solvent accessible (lower case), α-helix (red), β-strand (blue), hydrogen bond to main-chain amide group (bold) and hydrogen bond to main-chain carbonyl group (underlined). The frequency of amino acid substitutions is measured by each structural environment and averaged over all homologous protein families. Summing all 64 substitution tables leads to an environment-independent matrix such as PAM[125] (Point Accepted Mutation) or BLOSUM[126] (Block Substitution Matrix). Hence, ESSTs divide conventional substitution tables into 64 matrices, which differ by the local tertiary environment of amino acids in protein 3D structures. Ulla[127] is a programme that generates ESSTs from a set of structure alignments, annotated in various structural and functional environments for amino acid residues.

amino acid substitutions during evolution, although it should be noted that thermodynamically stable proteins are much more tolerant to mutations[56,57]. Based on the clustering pattern of 64 environments and other evidence mentioned earlier, it is clear that solvent accessibility is the primary structural constraint on amino acid substitutions and mutation rates during protein evolution.

Influence of hydrogen bonds on amino acid substitutions. Each of the three major clusters discussed above is further divided by the presence or absence of hydrogen bonds from side chains to main-chain NH groups (shown as the second concentric ring in FIG. 1). Hence, in either solvent-accessible or -inaccessible environments, the establishment of hydrogen bonds from side chains to main-chain NH groups restricts the substitution of amino acids, regardless of the local secondary structure. Interestingly, secondary structure (third concentric ring in FIG. 1), defined as α-helix, extended β-strand, positive φ torsion angle or coil, leads to the formation of further clusters within each of the clusters defined by the main-chain NH groups.

Amino acids with side-chain hydrogen bonds to main-chain CO groups (outermost concentric ring in FIG. 1) are grouped together in the secondary structure

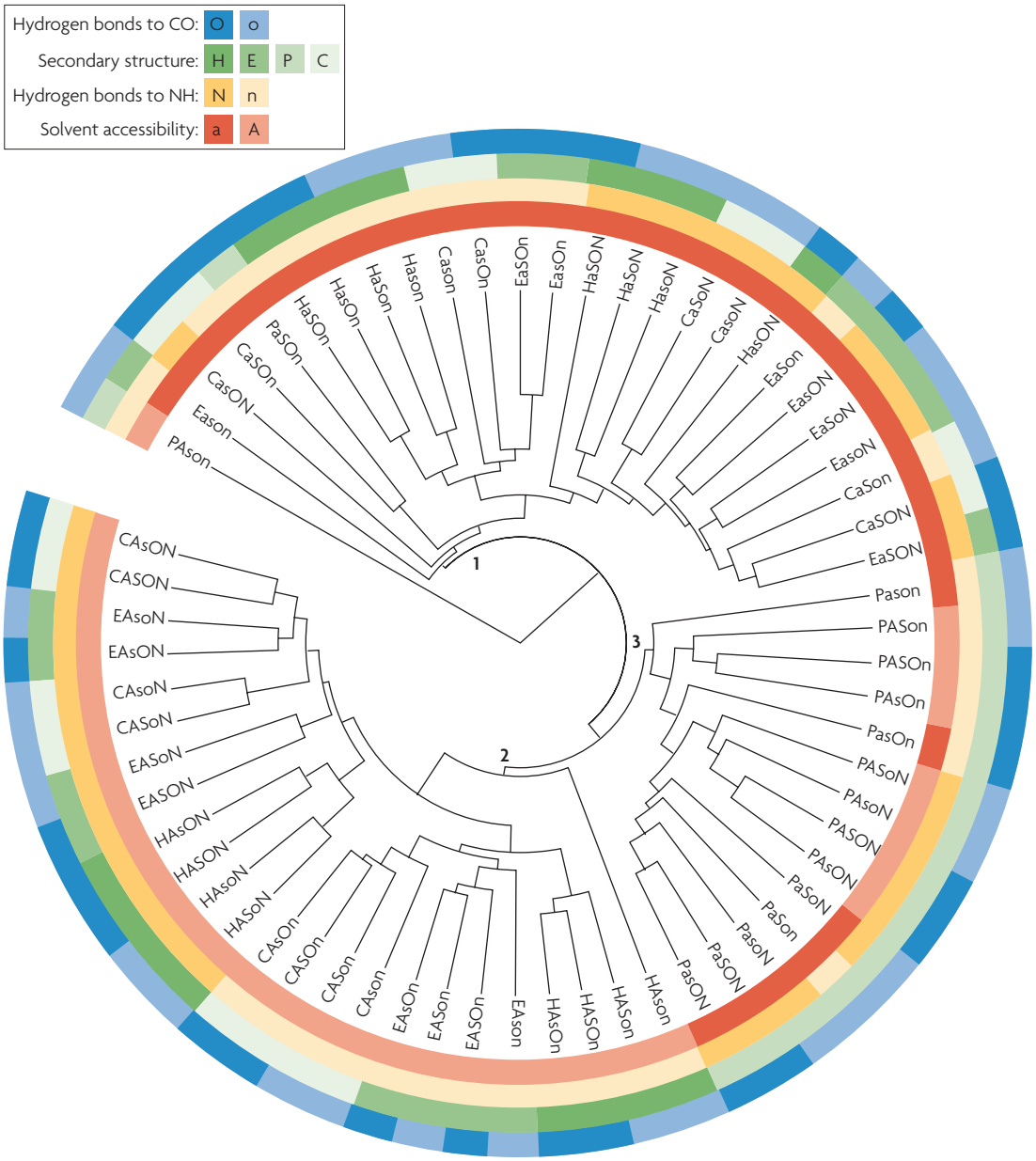


Figure 1 | **Results of hierarchical clustering of 64 environments.** Trees are constructed on the basis of the 64 × 64 distance matrix. Environments are shown using five-letter code representation: the first letter defines the secondary structure (α-helix (H), β-strand (E), positive φ main-chain torsion angle (P) and coil (C)), the second defines solvent accessibility (accessible (A) and inaccessible (a)) and the remaining three letters define the existence (upper case) or absence (lower case) of hydrogen bonds from a side chain to another side chain (S and s, third letter), to a main-chain carbonyl group (O and o, fourth letter) and to a main-chain amide group (N and n, fifth letter) (see also BOX 3 for details). Three major clusters are numbered as 1, 2 and 3 on the nodes from which they branch. Around the tree there are four concentric rings, each of which represents a particular structural parameter: the first ring represents solvent accessibility, the second ring represents the existence or absence of hydrogen bonds from a side chain to a main-chain amide group, the third ring represents the type of secondary structure and the fourth ring represents the existence or absence of hydrogen bonds from a side chain to a main-chain carbonyl group. The 4 concentric rings highlight the hierarchical clustering of the 64 environments by showing which amino acid substitution matrices are similar and which local environments are the major determinants of the substitution patterns. The trees were drawn using iTOL[128].

cluster, but the clustering pattern is weaker than that of main-chain NH groups. This suggests that the different types of hydrogen bonds have hierarchical effects on the substitution patterns of amino acids: hydrogen bonds between side chains and main-chain NH groups are most influential, followed by hydrogen bonds between main chains and main chains, and then side chains and main-chain CO groups. When the effects of solvent accessibility and then both solvent accessibility and the type of secondary structure are averaged, the clustering

retains the same order of hierarchy (see Supplementary information S1a,b (figure)). It is evident that there is a hierarchy in the influence of the eight types of hydrogen bonds from side chains on amino acid substitutions in homologous proteins (see Supplementary information S1c,d (figure)).

Positive φ torsion angles constrain protein evolution. In FIG. 1, matrices for the 64 environments with a positive φ torsion angle constitute a distinct cluster, whereas other elements of secondary structure are divided by solvent accessibility. A positive φ torsion angle can be accommodated by a Gly, which has no side chain, but for most other L-amino acids it leads to disallowed interactions between side-chain and main-chain atoms. However, for L-amino acids such as Asp or Asn, interactions between the side-chain CO group and the CO of the main-chain peptide bond can stabilize a positive φ angle conformation[58]. Indeed, Gly represents 63% of total amino acids that have a positive φ torsion angle, followed by Asn (8%) and Asp (5%) (data from ESSTs). In addition, in a positive φ angle class, solvent-accessible amino acids occur five times as frequently as inaccessible residues, whereas the average ratio of accessible to inaccessible residues is less than or equal to 2.2 for all classes of secondary structure. Hence, the predominance of Gly and polar residues in the set of amino acids with a positive φ torsion angle makes a distinct substitution pattern and eventually a distinct cluster.

The frequency of occurrence of local environments. Analysis of representative structures[59] of protein families shows that ~80% of all amino acids belong to 1 of 11 (out of 64) local environments (see Supplementary information S2 (table)). However, none of these 11 local environments includes any hydrogen bonds from side chains to main-chain NH groups, as expected from the observation that 68.6% of amino acids are non-polar and therefore cannot form hydrogen bonds with their side chains. Only 8.5% of amino acids have a side chain with a proton acceptor group and can therefore make hydrogen bonds from their side chains to main-chain NH groups, the second most important local environmental determinant of substitutions after solvent accessibility (see Supplementary information S3 (table)). The 8.5% of amino acids include Asp, Ser, Asn, Thr, Glu, Gln, Tyr, Met, Cys and His, and among them only Asp, Asn and Ser are over-represented compared with their background propensities in the protein data set. This shows that the distribution of amino acids taking part in hydrogen bonding from side chains to main chains follows the power law distribution — only a small proportion of amino acids have an important role in the substitution pattern.

We have shown that the degree of amino acid conservation is most affected by solvent accessibility, followed by the presence of hydrogen bonds from side chains to main chains and between main chains. However, there are other types of non-conventional interactions that are highly conserved and have important roles in protein structures and binding regions. Their importance is discussed in terms of protein stability later in this Review. A further consideration is the extent to which the local environment is conserved in homologous families and therefore can provide constraints on amino acid substitutions. Analyses of protein families and superfamilies show that the most crucial packing arrangements of individual side chains begin to differ when two proteins have less than 30% sequence identity. This is due to relative movements of equivalent secondary structural elements. However, some crucial hydrogen-bonding interactions are retained at much greater levels of sequence divergence.

Satisfaction of hydrogen bonds

Burying main chains and side chains in the interior of the protein removes them from the solvent and, through the hydrophobic effect, contributes much to the stability of the folded state of a protein. However, it is now clear that a comparable contribution to the stability of the folded protein is made by hydrogen bonding either within α-helices and β-sheets or through side chains forming hydrogen bonds with the unsatisfied NH and CO groups, as noted above. Indeed, the hydrogen-bonded side-chain groups occupy smaller volumes than the same groups when not hydrogen bonded. This leads to increased packing density and stronger van der Waals interactions in a protein[60], thus making a large, favourable contribution to protein stability and thereby to evolution[61].

Many side chains can make more than one hydrogen bond by acting as both proton donor and acceptor. Surveys of hydrogen bonds in sets of high-resolution protein structures have revealed that the polar atoms of a protein rarely fail to form hydrogen bonds and that they contribute to a hydrogen bond network that stabilizes the protein structure[34,62,63]. However, most studies that have looked at the satisfaction of hydrogen bonding potential in proteins have focused on main-chain interactions and have grouped side-chain interactions rather than treating each amino acid side chain separately[62,63]. Recently, an analysis of the hydrogen bonding potential of polar side chains in protein families has been described[64]. Unlike previous studies of hydrogen bonds in proteins, this study estimated the conservation of these polar residues in order to identify relationships that exist between residue conservation and satisfaction of hydrogen bond potential. Analysis of the sequence variability of buried amino acid residues in protein families shows that buried polar side chains, for which the hydrogen bond capacity is satisfied (that is, they form the full number of hydrogen bonds that they are capable of), are the most conserved amino acid residues in proteins. Buried and satisfied polar side chains are more conserved than non-polar residues and buried polar side chains that are unsatisfied or that do not form any hydrogen bonds.

Distinguishing the hydrogen-bonded state of a polar residue's side chain in terms of hydrogen bond satisfaction explains the observed conservation of these polar residues, particularly when the polar residue is buried. Where a polar residue is buried and satisfied in terms of

van der Waals interaction
A weak electrostatic interaction that is formed by the fluctuating electron clouds of two atoms.

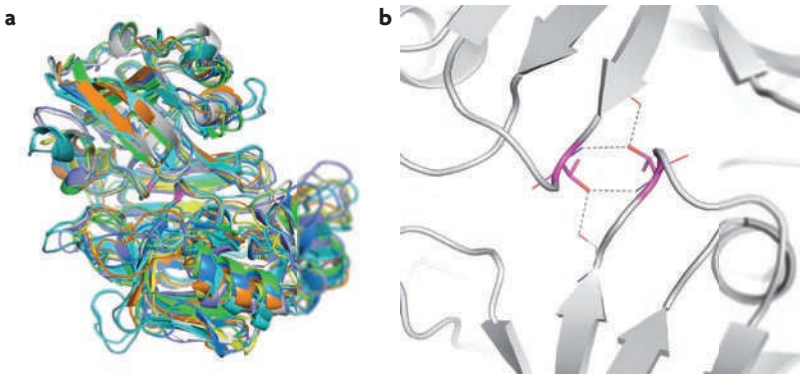


Figure 2 | **Structural constraints on pepsin-like aspartic proteinases. a** | Superimposed cartoon of eight members of the pepsin-like aspartic proteinase family that have two conserved buried Thr residues in topologically equivalent positions (shown in magenta), showing that hydrogen bonding interactions and the architectures that they stabilize are conserved in evolution. **b** | The two conserved Thr residues in a representative pepsin-like aspartic proteinase family member (Protein Data Bank code 3app). Each Thr forms two hydrogen bonds (shown as grey dashed lines) to main-chain atoms. These residues and the interactions that they form are conserved across the family, implying that the side-chain to main-chain interactions have an important role in the main-chain architecture of these proteins; in fact, the hydrogen bonds formed between the Thr residues and the main chain help to hold the two domains together. **c** | Selected regions of a multiple sequence alignment of the aspartic proteinases with two conserved DTG motifs (highlighted by black stars). The local structural environment of each residue in the alignment is indicated using IOY annotation[124]: solvent inaccessible (uppercase), solvent accessible (lower case), α-helix (red), β-strand (blue), hydrogen bond to side chain (overlined), hydrogen bond to main-chain amide group (bold), hydrogen bond to main-chain carbonyl group (underlined), disulphide bond (cedilla) and positive φ main-chain torsion angle (italic). Conserved α-helices and β-strands are indicated by 'a' and 'b' respectively. All protein structure images were produced using PyMOL.

side-chain hydrogen bonding, it is likely to have been conserved during evolution because it stabilizes the protein structure. Conversely, this same evolutionary pressure for conservation is not exerted on buried polar residues that are not hydrogen bonded or that are unsatisfied. Therefore, satisfaction of the hydrogen bonding potential of polar side chains is a key constraint in protein evolution.

Stabilization of protein architecture

So what kind of protein structures do these buried polar residues maintain? Most analyses of the stabilizing roles of polar side chains on the backbones of protein structures have focused on a particular secondary structural context[42,65–67]. One such study[68] analysed side-chain to side-chain and side-chain to main-chain interactions that were classified according to the position of the atom groups relative to the amino and carboxyl termini of α-helices, β-strands and coils. This and other analyses showed that 'capping' residues such as Glu or Asp interact with α-helix dipoles, which are formed by the aggregate effect of individual dipoles from all of the peptide groups in an α-helix and result in a partial positive charge at the α-helix N terminus and a partial negative charge at the C terminus[69]. Four- and five-residue motifs that begin with a Ser or Thr (ST motif)[37] or an Asp or Asn (Asx motif)[38] were identified. These motifs form hydrogen bonds from their respective polar side chains to the main-chain atoms of amino acids near the α-helix C terminus. The motifs help to stabilize protein structure when they occur at α-helix N termini but also commonly form independent ST β-turns or Asx β-turns or feature within β-bulge loops.

The key role that stabilizing hydrogen bond interactions have in maintaining protein structure is further demonstrated by an example that recurs in proteins: a highly conserved Tyr in the Tyr corner motif of immunoglobulin-like β-sandwich proteins is important for maintaining protein stability[67]. This is one of the many identified examples of recurring patterns that involve hydrogen bond interactions.

Analysis of the HOMSTRAD database[70] shows that out of a total of 142 protein families that have 5 members or more, 66 have entirely conserved buried polar residues and these equivalent residues form hydrogen bonds through their side chains to a main-chain atom in each structure. FIGURE 2 shows one such example of conservation of sequence and local structural environment for the aspartic proteinase family. The conservation of these side-chain to main-chain interactions implies that main-chain architecture is a crucial constraint on the evolution of proteins and that the interactions are retained as an essential part of the protein fold. Indeed, in this case it has been recognized that these hydrogen bonds contribute to holding together two domains, which seem to have evolved from identical subunits in an ancestral protein and are now retained in the dimeric retroviral proteinases, such as that from HIV.

What then can be said in more general terms about the architectures in which these interactions have such crucial roles? We now show that apart from capping local secondary structures, they often span elements of the secondary structure, in a way that is reminiscent of the roles of joists, braces or struts that span pillars and posts, and at other times support complex loop structures, like trusses that support the roofs of buildings.

Side chains spanning secondary structures. Typical examples of side chains that span secondary structures are provided by Asp residues, which frequently span α-helical N termini by forming hydrogen bonds to the N-terminal main-chain NH groups[36,71,72] of an adjacent α-helix. Such roles for Asp residues on α-helices provide strong constraints on their substitution by other amino acids. Arg residues have similar roles, spanning the C termini of α-helices. Thus, a buried Arg that is conserved in all seven members of the ribulose bisphosphate carboxylase family is always found at an equivalent position in an α-helix C terminus and is observed to span two helices, forming hydrogen bonds to the C terminus of the adjacent α-helix (FIG. 3a).

Tyr corner motif
A motif that involves a conserved Tyr within Greek key proteins forming a hydrogen bond with the local protein backbone in an adjacent loop.

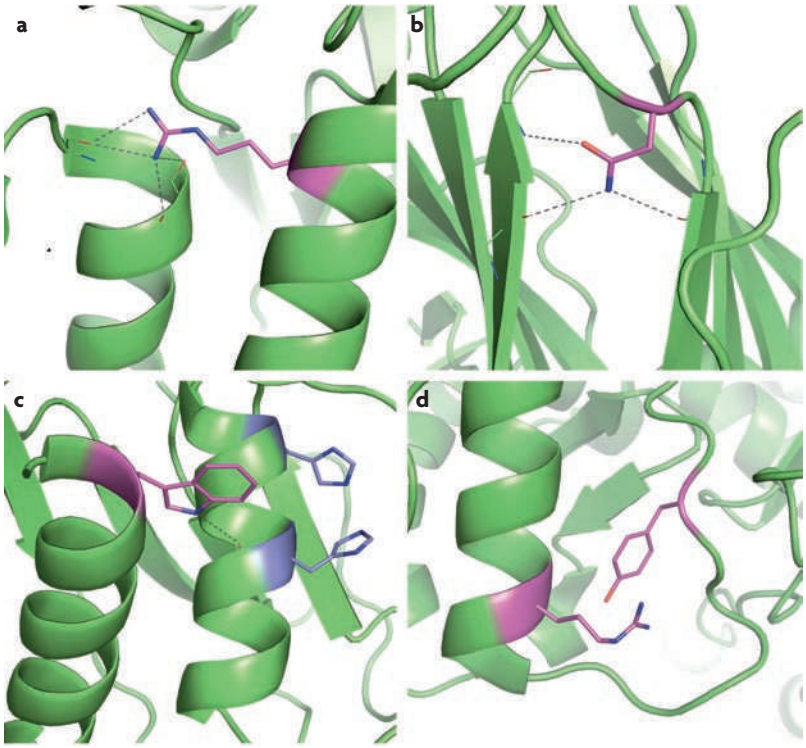


Figure 3 | **Conserved polar residues that span secondary structures. a** | A buried Arg at an α-helix carboxyl terminus in ribulose bisphosphate carboxylases (Protein Data Bank (PDB) code 1gk8) forming hydrogen bonds to another α-helix C terminus. **b** | A buried Asn spans β-strands in a β-barrel, forming a hydrogen bond with the main chain of another β-strand in the picornavirus coat protein family (PDB code 1tme). **c** | A Tyr in the matrix metalloproteinase family that spans α-helices, forming a hydrogen bond to a main-chain group in a second (distorted) α-helix that contains two active site His residues on the opposite face. **d** | An Arg in the glucocorticoid receptor family that forms a cation–π interaction with a Tyr residue (PDB code 1m2z). Representative structures were chosen for each family based on resolution; residues are coloured by atom type with buried polar residues shown in magenta. Hydrogen bonds are shown in grey.

Conserved side chains are also found spanning β-strands. This is often because main-chain atoms in β-strands are not satisfied by internal β-sheet hydrogen bonds and require side chains to satisfy their hydrogen bonding potential. This is frequently the case for edge strands (β-strands with one or no hydrogen bonding partner strand) or staggered β-strands, for example those in β-barrel structures. For instance, an entirely conserved and buried Asn residue in the picornavirus coat protein family forms hydrogen bonds with main-chain atoms in adjacent edge strands, providing a mechanism to satisfy the hydrogen bonding potential of these main-chain atoms (FIG. 3b).

Distortions in α-helices also lead to constraints on the substitution of buried polar residues. For example, in the matrix metalloproteinase family a buried Tyr hydrogen bonds to main-chain atoms in a distorted α-helix, probably helping to stabilize the active site His residues in a conformation that is necessary for catalysis (FIG. 3c).

Other weak non-covalent interactions such as aromatic–aromatic[73,74], amino–aromatic[75,76] and cation–π interactions[77] also provide a mechanism for stabilizing protein structure, and therefore lead to additional constraints on amino acid substitutions during divergent protein evolution. An interesting example is found in the glucocorticoid receptor family, in which a conserved Arg forms a cation–π interaction with a conserved and buried Tyr (FIG. 3d). By means of structural, phylogenetic and functional analyses, it was shown that mutation from Tyr to Arg at position 27 in an ancestral protein of the glucocorticoid receptor must have increased stability on a crucial part of the receptor[78]. The authors postulate that although this mutation had no immediate consequence, it created a permissive sequence environment for substitutions that, millions of years later, remodelled the protein and yielded a new function.

All of the conserved buried polar residues shown in FIG. 3a–d have roles that in many ways are analogous to those of struts or joists in buildings. We conclude that it is not only the local environment but also its role in the context of the overall architecture and function that places the constraints on amino acid substitutions.

Side chains supporting coils and turns. In regions of extensive non-regular secondary structure, amino acid residues are often unable to form intra-main-chain hydrogen bonds. Such structures are often supported by polar side chains from secondary structure elements. Examples occur with twisted lone β-strands, short α-helices and complex loop structures. All provide architectural requirements for constraints on local environments.

For example, the Ca^{2+}-binding, parvalbumin-like proteins have an entirely conserved and buried Asp that forms hydrogen bonds to a coil region (FIG. 4a). A similar interaction is observed in the interleukin-1β-like growth factor family (although in this case it emerges from a β-strand rather than an α-helix), in which a conserved and buried Ser forms hydrogen bonds to main-chain atoms in type I and type IV β-turns (FIG. 4b). The conserved buried polar residues in these two examples help to stabilize regions of coil, often in elaborate loop structures that form extended turns and arches. Previous analyses of intra-coil side-chain to main-chain hydrogen bonds revealed that Asp, Ser, Asn and Thr are the polar residues that most commonly form this type of interaction, with 80% of these cases being at solvent-exposed sites[68].

The alcohol dehydrogenase family has a conserved buried Arg residue that forms hydrogen bonds to polyproline α-helices (FIG. 4c). In fact, Arg is the most common polar residue to form hydrogen bonds to main-chain atoms of polyproline-type α-helices[43], in which intra-chain hydrogen bonds cannot form owing to the extended nature of the chains and in which the three-fold screw rotation symmetry prevents extensive super-secondary interactions of the kind found in β-sheets. Instead, side-chain to main-chain hydrogen bonds contribute to main-chain atom satisfaction and polyproline stability.

All of the conserved buried polar residues shown in FIG. 4a–c involve multiple side-chain to main-chain interactions that often form structures resembling the trusses of roof supports and bridges. Buried polar amino acids

Cation–π interaction
A non-covalent interaction between an aromatic side chain and a cationic side chain.

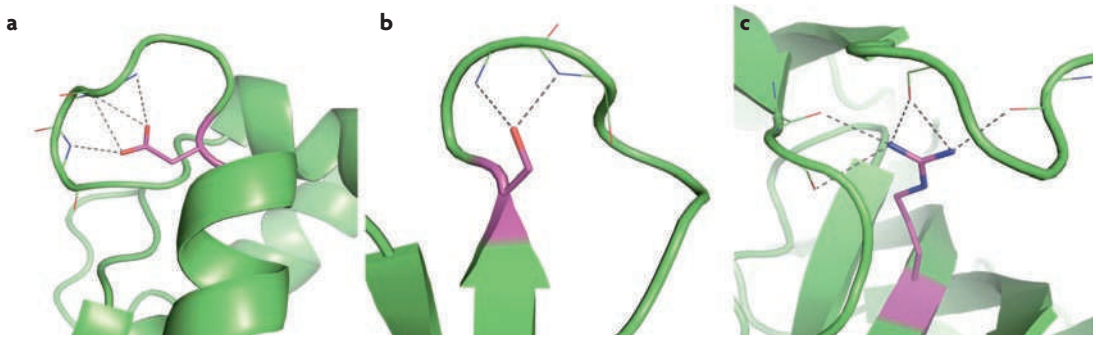


Figure 4 | **Conserved polar residues that support loops. a** | An Asp forming hydrogen bonds to a coil region in the Ca^{2+}-binding, parvalbumin-like proteins (Protein Data Bank (PDB) code 5pal). **b** | A Ser forming hydrogen bonds to main-chain atoms in type I and type IV β-turns in interleukin-1β-like growth factor family proteins (PDB code 2fgf). **c** | An Arg forming hydrogen bonds to polyproline α-helices in the alcohol dehydrogenases (polyproline interaction on the right) (PDB code 2ohxa). Representative structures were chosen for each family based on resolution; residues are coloured by atom type with buried polar residues shown in magenta. Hydrogen bonds are shown in grey.

can provide a mechanism of pinning loops in place when main-chain to main-chain interactions cannot suffice. Conservation of these residues and the interactions that they form implies that they are important for maintaining protein structure and therefore can provide strong constraints on amino acid substitutions.

Evolutionary pressure on fast folding

Residues that ensure fast and correct protein folding also ensure correct function and this also leads to constraints on the evolution of proteins. Folding simulations and sequence design have been used to develop a method for determining the folding nucleus of a protein with known structure. This method has been applied to chymotrypsin inhibitor 2 (REF. 79). The predicted set of folding nucleus residues matched those identified by kinetic studies[80], with a clear qualitative correlation being observed between site conservation and φ-values for folding. This indicates the importance of a given residue to the structure of the folding nucleus by providing a quantitative measure of the extent to which a residue participates in native-like interactions during the rate-limiting step in folding. The study implies that residues that are involved in the folding nucleus, and hence are important for forming the native protein structure, constrain amino acid substitutions.

Mirny *et al.* developed the 'conservatism of conservatism' principle for analysing evolutionary signals that are specific to a given fold; that is, they identified conserved amino acid positions in families of proteins that are structurally related to one another (but not related by sequence)[81]. This approach identified residues that belong to the folding nucleus of chemotaxis protein CheY[81]. Subsequent application to five of the most common protein folds demonstrated that evolutionary pressure towards fast folding and function can also lead to higher conservation of residues than expected from solvent accessibility[82]. However, other researchers are not all in agreement about fast folding constraining the evolution of proteins; for example, Baker and co-workers did not observe a correlation between conservation and experimentally measured φ-values[83]. Nevertheless, they did observe a significant correlation between the contribution of individual sequence positions and the transition state structure among homologous proteins, indicating that the structure of the folding transition state ensemble seems to be more highly conserved than the specific interactions that stabilize it[83].

Further studies have indicated that poorly and highly conserved residues are equally likely to participate in the protein-folding nucleus, igniting further controversy on the notion of folding nucleus conservation[84,85]. However, these later studies confirmed that the folding nucleus of CheY is significantly conserved, although this was the exception in the protein data sets studied and is perhaps due to extraordinarily tight packing of the folding nucleus in CheY[76]. The folding nuclei of some proteins contain non-native interactions in the transition state that, when weakened, slow folding down but do not change the protein stability. This is illustrated by a universally conserved Ile in the SH3 domain, which is kinetically but not thermodynamically important in the SH3 domain-containing protein Tyr kinase Src (REF. 86). Therefore, evolutionary constraints on protein structure act both to maintain protein architecture and to maintain correct (and fast) folding.

Maintenance of function

All of the constraints related to maintenance of tertiary structure are ultimately functional. However, many functions are mediated through quaternary interactions of proteins with other macromolecules in assemblies or with substrates, ligands or allosteric regulators. The effects of these constraints are felt some distance away from the interaction site but they tend to have an increasing influence nearer to the recognition site. To investigate this, the Euclidean distance was measured between every amino acid and the known functional residues and the degree of conservation was compared in terms of the proximity with functional residues[51]. The authors showed that the degree of residue conservation is significantly higher in residues that are near to the active site than in those that are far from it. Hence, geometrical distance from known active sites constitutes another

SH3 domain
(Src homology 3 domain). A small domain that is found in various intracellular or membrane-associated proteins and has a β-barrel fold.

Euclidean distance
A geometric distance between two point sets in the *n*-dimensional (or Euclidean) space.

constraint on amino acid substitutions in protein evolution and therefore can serve as an additional parameter to define the local structural environment in classifying amino acid substitution patterns.

The impact of various functional constraints — mainly defined in terms of interactions with other molecules such as substrates, ligands, nucleic acids and other proteins — on the conservation of amino acids in three-dimensional structures has been investigated. Functional residues were excluded (masked) from the sequence alignment, and the degree of residue conservation was measured by discarding the locations of functional residues from the calculation of substitution probabilities[59]. Several masking models were prepared by using various combinations of functional residues and were compared with the non-masking model, which includes functional residues in the calculation of substitution probabilities. The average probability of amino acid conservation for the non-masking model was ~1.36% higher than that of a masking model, although the difference was less distinct when enzyme active sites were omitted from masking[59]. Overall this shows that functional residues are under greater pressure to be conserved throughout the evolutionary process when they are crucially important to the activity of proteins and thus confer selective advantage to the organism.

Mutations that occur in functional residues lead to loss-of-function of the protein either by disrupting the native structure or by interfering in the interaction with other molecules. However, mutations are sometimes compensated by other mutations occurring in the interacting partner molecule or molecules, which is explained as co-adaptation or co-evolution of interacting protein pairs[87,88].

Conclusions

We have discussed how structural and functional features constrain the evolution of proteins, with both being driven by the maintenance of protein function. The identification of such constraints in protein families can be helpful for protein engineering experiments, such as designing enzymes with new functions or in the directed stabilization of protein conformations through site-directed mutagenesis. Understanding such features also allows the identification of members of a superfamily and often the prediction of functionally important interacting regions, so providing valuable annotation of genome sequences in terms of structure and function.

In this Review we have focused on constraints on the substitution of individual amino acids. Strong constraints arise from the conservation of structure, not only from maintenance of a hydrophobic core and secondary structure but also from buried, often charged hydrogen bonds. However, we have discussed how constraints also arise from interactions with other proteins; these are often components of interaction networks that are conserved throughout evolution[89], so that interacting proteins are under various constraints such as activity and lifetime[90–92]. Other factors can also be correlated with the rate of protein evolution. For example, expression level might be an important factor influencing evolutionary rate[93–95] as highly expressed proteins are constrained to have fewer mutations than rarer proteins to avoid the cost of misfolding effects. A proper understanding of the constraints on amino acid substitutions is an essential prerequisite to understanding protein evolution, but further insights will depend on integrated and multidisciplinary systems approaches[23,96].

1. Bajaj, M. & Blundell, T. Evolution and the tertiary structure of proteins. *Annu. Rev. Biophys. Bioeng.* **13**, 453–492 (1984).
2. Chothia, C. & Lesk, A. M. The relation between the divergence of sequence and structure in proteins. *EMBO J.* **5**, 823–826 (1986).
This paper quantifies the relationship between sequence variance and structural tolerance.
3. Kimura, M. Evolutionary rate at the molecular level. *Nature* **217**, 624–626 (1968).
The first paper to introduce the neutral theory of evolution.
4. Ohta, T. Slightly deleterious mutant substitutions in evolution. *Nature* **246**, 96–98 (1973).
Introduces the nearly neutral theory of molecular evolution, a modification of that detailed in reference 3.
5. Zuckerkandl, E. Evolutionary processes and evolutionary noise at the molecular level. I. Functional density in proteins. *J. Mol. Evol.* **7**, 167–183 (1976).
6. Zuckerkandl, E. Evolutionary processes and evolutionary noise at the molecular level. II. A selectionist model for random fixations in proteins. *J. Mol. Evol.* **7**, 269–311 (1976).
7. Fraser, H. B., Hirsh, A. E., Steinmetz, L. M., Scharfe, C. & Feldman, M. W. Evolutionary rate in the protein interaction network. *Science* **296**, 750–752 (2002).
8. Bloom, J. D. & Adami, C. Apparent dependence of protein evolutionary rate on number of interactions is linked to biases in protein-protein interactions data sets. *BMC Evol. Biol.* **3**, 21 (2003).
9. Jordan, I. K., Wolf, Y. I. & Koonin, E. V. No simple dependence between protein evolution rate and the number of protein–protein interactions: only the most prolific interactors tend to evolve slowly. *BMC Evol. Biol.* **3**, 1 (2003).
10. Orengo, C. A. & Thornton, J. M. Protein families and their evolution — a structural perspective. *Annu. Rev. Biochem.* **74**, 867–900 (2005).
11. Bullock, A. N. *et al.* Thermodynamic stability of wild-type and mutant p53 core domain. *Proc. Natl Acad. Sci. USA* **94**, 14338–14342 (1997).
An elegant study that applied techniques initially devised to study the biophysics of protein folding to mutations in the protein p53, demonstrating that most of these changes are destabilizing.
12. Canadillas, J. M. *et al.* Solution structure of p53 core domain: structural basis for its instability. *Proc. Natl Acad. Sci. USA* **103**, 2109–2114 (2006).
13. Friedler, A., Veprintsev, D. B., Hansson, L. O. & Fersht, A. R. Kinetic instability of p53 core domain mutants: implications for rescue by small molecules. *J. Biol. Chem.* **278**, 24108–24112 (2003).
14. Joerger, A. C., Allen, M. D. & Fersht, A. R. Crystal structure of a superstable mutant of human p53 core domain. Insights into the mechanism of rescuing oncogenic mutations. *J. Biol. Chem.* **279**, 1291–1296 (2004).
15. Nikolova, P. V., Henckel, J., Lane, D. P. & Fersht, A. R. Semirational design of active tumor suppressor p53 DNA binding domain with enhanced stability. *Proc. Natl Acad. Sci. USA* **95**, 14675–14680 (1998).
16. Wang, X., Minasov, G. & Shoichet, B. K. Evolution of an antibiotic resistance enzyme constrained by stability and activity trade-offs. *J. Mol. Biol.* **320**, 85–95 (2002).
17. Aharoni, A. The 'evolvability' of promiscuous protein functions. *Nature Genet.* **37**, 73–76 (2005).
An original study on the evolution of new protein functions that shows that the process is driven by mutations having little effect on native function but large effects on promiscuous function.
18. Aharoni, A. *et al.* Directed evolution of mammalian paraoxonases PON1 and PON3 for bacterial expression and catalytic specialization. *Proc. Natl Acad. Sci. USA* **101**, 482 (2004).
19. Andreeva, A. & Murzin, A. G. Evolution of protein fold in the presence of functional constraints. *Curr. Opin. Struct. Biol.* **16**, 399–408 (2006).
A review of the mechanisms by which a protein fold can evolve whilst maintaining the functional-site structure.
20. Caetano-Anollés, G., Wang, M., Caetano-Anollés, D. & Mittenthal, J. E. The origin, evolution and structure of the protein world. *Biochem. J.* **417**, 621–637 (2009).
21. Copley, R. R., Letunic, I. & Bork, P. Genome and protein evolution in eukaryotes. *Curr. Opin. Chem. Biol.* **6**, 39–45 (2002).
22. Kinch, L. N. & Grishin, N. V. Evolution of protein structures and functions. *Curr. Opin. Struct. Biol.* **12**, 400–408 (2002).
23. Pal, C., Papp, B. & Lercher, M. J. An integrated view of protein evolution. *Nature Rev. Genet.* **7**, 337–348 (2006).
A comprehensive review of various approaches to study protein evolution.
24. Koonin, E. V. Orthologs, paralogs, and evolutionary genomics. *Annu. Rev. Genet.* **39**, 309–338 (2005).
25. Hubbard, T. J. & Blundell, T. L. Comparison of solvent-inaccessible cores of homologous proteins: definitions useful for protein modelling. *Protein Eng.* **1**, 159–171 (1987).
26. Garnier, J., Osguthorpe, D. J. & Robson, B. Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins. *J. Mol. Biol.* **120**, 97–120 (1978).

27. Gibrat, J. F., Garnier, J. & Robson, B. Further developments of protein secondary structure prediction using information theory. New parameters and consideration of residue pairs. *J. Mol. Biol.* **198**, 425–443 (1987).
28. Levin, J. M., Robson, B. & Garnier, J. An algorithm for secondary structure determination in proteins based on sequence similarity. *FEBS Lett.* **205**, 303 (1986).
29. Pauling, L. & Corey, R. B. Configurations of polypeptide chains with favored orientations around single bonds: two new pleated sheets. *Proc. Natl Acad. Sci. USA* **37**, 729–740 (1951).
30. Pauling, L., Corey, R. B. & Branson, H. R. The structure of proteins; two hydrogen-bonded helical configurations of the polypeptide chain. *Proc. Natl Acad. Sci. USA* **37**, 205–211 (1951).
References 29 and 30 provided the first hint that regular secondary structure might form in folded proteins.
31. Hutchinson, E. G. & Thornton, J. M. A revised set of potentials for β-turn formation in proteins. *Protein Sci.* **3**, 2207–2216 (1994).
32. Sibanda, B. L., Blundell, T. L. & Thornton, J. M. Conformation of β-hairpins in protein structures. A systematic classification with applications to modelling by homology, electron density fitting and protein engineering. *J. Mol. Biol.* **206**, 759–777 (1989).
33. Wilmot, C. M. & Thornton, J. M. Analysis and prediction of the different types of β-turn in proteins. *J. Mol. Biol.* **203**, 221–232 (1988).
34. Baker, E. N. & Hubbard, R. E. Hydrogen bonding in globular proteins. *Prog. Biophys. Mol. Biol.* **44**, 97–179 (1984).
The first comprehensive survey of hydrogen bonds in high-resolution protein structures.
35. Presta, L. G. & Rose, G. D. Helix signals in proteins. *Science* **240**, 1632–1641 (1988).
36. Richardson, J. S. & Richardson, D. C. Amino acid preferences for specific locations at the ends of α helices. *Science* **240**, 1648–1652 (1988).
37. Wan, W. Y. & Milner-White, E. J. A recurring two-hydrogen-bond motif incorporating a serine or threonine residue is found both at α-helical N termini and in other situations. *J. Mol. Biol.* **286**, 1651–1662 (1999).
38. Wan, W. Y. & Milner-White, E. J. A natural grouping of motifs with an aspartate or asparagine residue forming two hydrogen bonds to residues ahead in sequence: their occurrence at α-helical N termini and in other situations. *J. Mol. Biol.* **286**, 1633–1649 (1999).
39. Chan, A. W. E., Hutchinson, E. G. & Thornton, J. M. Identification, classification, and analysis of β-bulges in proteins. *Protein Sci.* **2**, 1574–1590 (1993).
40. Richardson, J. S., Getzoff, E. D. & Richardson, D. C. The β bulge: a common small unit of nonrepetitive protein structure. *Proc. Natl Acad. Sci. USA* **75**, 2574–2578 (1978).
41. Barlow, D. J. & Thornton, J. M. Helix geometry in proteins. *J. Mol. Biol.* **201**, 601–619 (1988).
42. Eswar, N. & Ramakrishnan, C. Secondary structures without backbone: an analysis of backbone mimicry by polar side chains in protein structures. *Protein Eng.* **12**, 447–455 (1999).
43. Cubellis, M. V., Caillez, F., Blundell, T. L. & Lovell, S. C. Properties of polyproline II, a secondary structure element implicated in protein–protein interactions. *Proteins* **58**, 880–892 (2005).
44. Stapley, B. J. & Creamer, T. P. A survey of left-handed polyproline II helices. *Protein Sci.* **8**, 587–595 (1999).
45. Milner-White, E., Ross, B. M., Ismail, R., Belhadj-Mostefa, K. & Poet, R. One type of γ-turn, rather than the other gives rise to chain-reversal in proteins. *J. Mol. Biol.* **204**, 777–782 (1988).
46. Milner-White, E. J. β-bulges within loops as recurring features of protein structure. *Biochim. Biophys. Acta* **911**, 261–265 (1987).
47. Blundell, T. L. & Wood, S. P. Is the evolution of insulin Darwinian or due to selectively neutral mutation? *Nature* **257**, 197–203 (1975).
An early paper discussing the evolution of protein structure and interactions in terms of adaptive processes and neutral mutations.
48. Guharoy, M. & Chakrabarti, P. Conservation and relative importance of residues across protein–protein interfaces. *Proc. Natl Acad. Sci. USA* **102**, 15447–15452 (2005).
49. Kisters-Woike, B., Vangierdegom, C. & Mueller-Hill, B. On the conservation of protein sequences in evolution. *Trends Biochem. Sci.* **25**, 419–421 (2000).
50. Lichtarge, O., Bourne, H. R. & Cohen, F. E. Evolutionarily conserved Gαβγ binding surfaces support a model of the G protein-receptor complex. *Proc. Natl Acad. Sci. USA* **93**, 7507–7511 (1996).
51. Chelliah, V., Chen, L., Blundell, T. L. & Lovell, S. C. Distinguishing structural and functional restraints in evolution in order to identify interaction sites. *J. Mol. Biol.* **342**, 1487–1504 (2004).
52. Blundell, T. L. *et al.* in *Methods in Proteins Sequence Analysis* (eds Jornvall, H. Hoog, J.O. Gustavsson, A.M.) 373–385 (Birkhauser, Basel, 1991).
53. Overington, J., Johnson, M. S., Sali, A. & Blundell, T. L. Tertiary structural constraints on protein evolutionary diversity: templates, key residues and structure prediction. *Proc. Biol. Sci.* **241**, 132–145 (1990).
The first study to quantify structural restraints on amino acid substitutions between homologous proteins, identifying particular patterns of substitution.
54. Overington, J., Donnelly, D., Johnson, M. S., Sali, A. & Blundell, T. L. Environment-specific amino acid substitution tables: tertiary templates and prediction of protein folds. *Protein Sci.* **1**, 216–226 (1992).
55. Michener, C. D. & Sokal, R. R. A quantitative approach to a problem in classification. *Evolution* **11**, 130 (1957).
56. Bloom, J. D., Labthavikul, S. T., Otey, C. R. & Arnold, F. H. Protein stability promotes evolvability. *Proc. Natl Acad. Sci. USA* **103**, 5869–5874 (2006).
57. Bloom, J. D. *et al.* Thermodynamic prediction of protein neutrality. *Proc. Natl Acad. Sci. USA* **102**, 606–611 (2005).
58. Deane, C. M., Allen, F. H., Taylor, R. & Blundell, T. L. Carbonyl–carbonyl interactions stabilize the partially allowed Ramachandran conformations of asparagine and aspartic acid. *Protein Eng.* **12**, 1025–1028 (1999).
59. Gong, S. & Blundell, T. L. Discarding functional residues from the substitution table improves predictions of active sites within three-dimensional structures. *PLoS Comput. Biol.* **4**, e1000179 (2008).
60. Schell, D., Tsai, J., Scholtz, J. M. & Pace, C. N. Hydrogen bonding increases packing density in the protein interior. *Proteins* **63**, 278–282 (2006).
61. Pace, C. N. Polar group burial contributes more to protein stability than nonpolar group burial. *Biochemistry* **16**, 310–313 (2001).
62. Fleming, P. J. & Rose, G. D. Do all backbone polar groups in proteins form hydrogen bonds? *Protein Sci.* **14**, 1911–1917 (2005).
63. McDonald, I. K. & Thornton, J. M. Satisfying hydrogen bonding potential in proteins. *J. Mol. Biol.* **238**, 777–793 (1994).
64. Worth, C. L. & Blundell, T. L. Satisfaction of hydrogen-bonding potential influences the conservation of polar sidechains. *Proteins* **75**, 413–429 (2009).
65. Eswar, N. & Ramakrishnan, C. Deterministic features of side-chain main-chain hydrogen bonds in globular protein structures. *Protein Eng.* **13**, 227–238 (2000).
66. Vijayakumar, M., Qian, H. & Zhou, H. X. Hydrogen bonds between short polar side chains and peptide backbone: prevalence in proteins and effects on helix-forming propensities. *Proteins* **34**, 497–507 (1999).
67. Hamill, S. J., Cota, E., Chothia, C. & Clarke, J. Conservation of folding and stability within a protein family: the tyrosine corner as an evolutionary cul-de-sac. *J. Mol. Biol.* **295**, 641–649 (2000).
68. Bordo, D. & Argos, P. The role of side-chain hydrogen bonds in the formation and stabilization of secondary structure in soluble proteins. *J. Mol. Biol.* **243**, 504–519 (1994).
69. Nicholson, H., Anderson, D. E., Dao-pin, S. & Matthews, B. W. Analysis of the interaction between charged side chains and the α-helix dipole using designed thermostable mutants of phage T4 lysozyme. *Biochemistry* **30**, 9816–9828 (1991).
70. Mizuguchi, K., Deane, C. M., Blundell, T. L. & Overington, J. P. HOMSTRAD: a database of protein structure alignments for homologous families. *Protein Sci.* **7**, 2469–2471 (1998).
71. Harper, E. T. & Rose, G. D. Helix stop signals in proteins and peptides: the capping box. *Biochemistry* **32**, 7605–7609 (1993).
72. Serrano, L., Sancho, J., Hirshberg, M. & Fersht, A. R. α-Helix stability in proteins. I. Empirical correlations concerning substitution of side-chains at the N and C-caps and the replacement of alanine by glycine or serine at solvent-exposed surfaces. *J. Mol. Biol.* **227**, 544–559 (1992).
73. Burley, S. K. & Petsko, G. A. Aromatic–aromatic interaction — a mechanism of protein-structure stabilization. *Science* **229**, 23–28 (1985).
74. Hunter, C. A., Singh, J. & Thornton, J. M. Pi–Pi-interactions — the geometry and energetics of phenylalanine phenylalanine interactions in proteins. *J. Mol. Biol.* **218**, 837–846 (1991).
75. Burley, S. K. & Petsko, G. A. Amino-aromatic interactions in proteins. *FEBS Lett.* **203**, 139–143 (1986).
76. Mitchell, J. B. O., Nandi, C. L., Mcdonald, I. K., Thornton, J. M. & Price, S. L. Amino/aromatic interactions in proteins — is the evidence stacked against hydrogen-bonding. *J. Mol. Biol.* **239**, 315–331 (1994).
77. Gallivan, J. P. & Dougherty, D. A. Cation–π interactions in structural biology. *Proc. Natl Acad. Sci. USA* **96**, 9459–9464 (1999).
78. Ortlund, E. A., Bridgham, J. T., Redinbo, M. R. & Thornton, J. W. Crystal structure of an ancient protein: evolution by conformational epistasis. *Science* **317**, 1544–1548 (2007).
79. Shakhnovich, E., Abkevich, V. & Ptitsyn, O. Conserved residues and the mechanism of protein folding. *Nature* **379**, 96–98 (1996).
The presentation of a novel computational method for identifying the residues that form the folding nucleus of a protein.
80. Itzhaki, L. S., Otzen, D. E. & Fersht, A. R. The structure of the transition state for folding of chymotrypsin inhibitor 2 analysed by protein engineering methods: evidence for a nucleation-condensation mechanism for protein folding. *J. Mol. Biol.* **254**, 260–288 (1995).
Introduced the nucleation–condensation model of protein folding from experimental work in chymotrypsin inhibitor 2.
81. Mirny, L. A., Abkevich, V. I. & Shakhnovich, E. I. How evolution makes proteins fold quickly. *Proc. Natl Acad. Sci. USA* **95**, 4976–4981 (1998).
82. Mirny, L. A. & Shakhnovich, E. I. Universally conserved positions in protein folds: reading evolutionary signals about stability, folding kinetics and function. *J. Mol. Biol.* **291**, 177–196 (1999).
83. Plaxco, K. W. *et al.* Evolutionary conservation in protein folding kinetics. *J. Mol. Biol.* **298**, 303 (2000).
84. Larson, S. M., Ruczinski, I., Davidson, A. R., Baker, D. & Plaxco, K. W. Residues participating in the protein folding nucleus do not exhibit preferential evolutionary conservation. *J. Mol. Biol.* **316**, 225–233 (2002).
85. Tseng, Y. Y. & Liang, J. Are residues in a protein folding nucleus evolutionarily conserved? *J. Mol. Biol.* **335**, 869–880 (2004).
86. Li, L., Mirny, L. A. & Shakhnovich, E. I. Kinetics, thermodynamics and evolution of non-native interactions in a protein folding nucleus. *Nature Struct. Biol.* **7**, 336–342 (2000).
87. Kim, W. K., Bolser, D. M. & Park, J. H. Large-scale co-evolution analysis of protein structural interlogues using the global protein structural interactome map (PSIMAP). *Bioinformatics* **20**, 1138–1150 (2004).
88. Pazos, F. & Valencia, A. Protein co-evolution, co-adaptation and interactions. *EMBO J.* **27**, 2648–2655 (2008).
89. Park, J. & Bolser, D. Conservation of protein interaction network in evolution. *Genome Inform.* **12**, 135–140 (2001).
90. Batada, N. N., Hurst, L. D. & Tyers, M. Evolutionary and physiological importance of hub proteins. *PLoS Comput. Biol.* **2**, e88 (2006).
91. Pal, C., Papp, B. & Hurst, L. D. Genomic function: rate of evolution and gene dispensability. *Nature* **421**, 496–497 (2003).
92. Wall, D. P. *et al.* Functional genomic analysis of the rates of protein evolution. *Proc. Natl Acad. Sci. USA* **102**, 5483–5488 (2005).
93. Choi, J. K., Kim, S. C., Seo, J., Kim, S. & Bhak, J. Impact of transcriptional properties on essentiality and evolutionary rate. *Genetics* **175**, 199–206 (2007).
94. Drummond, D. A., Bloom, J. D., Adami, C., Wilke, C. O. & Arnold, F. H. Why highly expressed proteins evolve slowly. *Proc. Natl Acad. Sci. USA* **102**, 14338–14343 (2005).
This paper suggests that the expression level of a protein is related to the demand for exact folding.
95. Drummond, D. A., Raval, A. & Wilke, C. O. A single determinant dominates the rate of yeast protein evolution. *Mol. Biol. Evol.* **23**, 327–337 (2006).
96. Zeldovich, K. B. & Shakhnovich, E. I. Understanding protein evolution: from protein physics to Darwinian selection. *Annu. Rev. Phys. Chem.* **59**, 105–127 (2008).
97. Akashi, H. Gene expression and molecular evolution. *Curr. Opin. Genet. Dev.* **11**, 660–666 (2001).

98. Drummond, D. A. & Wilke, C. O. Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution. *Cell* **134**, 341–352 (2008).
99. Hamill, S. J., Steward, A. & Clarke, J. The folding of an immunoglobulin-like Greek key protein is defined by a common-core nucleus and regions constrained by topology. *J. Mol. Biol.* **297**, 165 (2000).
100. Chiti, F. & Dobson, C. M. Protein misfolding, functional amyloid, and human disease. *Annu. Rev. Biochem.* **75**, 333–366 (2006).
101. Hamada, D. *et al.* Competition between folding, native-state dimerisation and amyloid aggregation in β-lactoglobulin. *J. Mol. Biol.* **386**, 878–890 (2009).
102. Goldberg, A. L. Protein degradation and protection against misfolded or damaged proteins. *Nature* **426**, 895–899 (2003).
103. Wolffe, A. P. & Matzke, M. A. Epigenetics: regulation through repression. *Science* **286**, 481–486 (1999).
104. Murzin, A. G., Brenner, S. E., Hubbard, T. & Chothia, C. SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* **247**, 536–540 (1995).
Details the first protein hierarchical classification scheme.
105. Orengo, C. A. *et al.* CATH–a hierarchic classification of protein domain structures. *Structure* **5**, 1093–1108 (1997).
106. Bhaduri, A., Pugalenthi, G. & Sowdhamini, R. PASS2: an automated database of protein alignments organised as structural superfamilies. *BMC Bioinformatics* **5**, 35 (2004).
107. Worth, C. L. *et al.* A structural bioinformatics approach to the analysis of nonsynonymous single nucleotide polymorphisms (nsSNPs) and their relation to disease. *J. Bioinform. Comput. Biol.* **5**, 1297–1318 (2007).
108. Holm, L., Kaariainen, S., Rosenstrom, P. & Schenkel, A. Searching protein structure databases with DaliLite v.3. *Bioinformatics* **24**, 2780 (2008).
109. Shindyalov, I. N. & Bourne, P. E. Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. *Protein Eng.* **11**, 739–747 (1998).
110. Marchler-Bauer, A. *et al.* MMDB: Entrez's 3D structure database. *Nucleic Acids Res.* **27**, 240–243 (1999).
111. Finn, R. D. *et al.* The Pfam protein families database. *Nucleic Acids Res.* **36**, D281–D288 (2008).
112. Hunter, S. *et al.* InterPro: the integrative protein signature database. *Nucleic Acids Res.* **37**, D211–D215 (2009).
113. Hulo, N. *et al.* The PROSITE database. *Nucleic Acids Res.* **34**, D227–D230 (2006).
114. Attwood, T. K. *et al.* PRINTS and its automatic supplement, prePRINTS. *Nucleic Acids Res.* **31**, 400–402 (2003).
115. Servant, F. *et al.* ProDom: automated clustering of homologous domains. *Brief. Bioinformatics* **3**, 246–251 (2002).
116. Schultz, J., Milpetz, F., Bork, P. & Ponting, C. P. SMART, a simple modular architecture research tool: identification of signaling domains. *Proc. Natl Acad. Sci. USA* **95**, 5857–5864 (1998).
117. Haft, D. H., Selengut, J. D. & White, O. The TIGRFAMs database of protein families. *Nucleic Acids Res.* **31**, 371–373 (2003).
118. Buchan, D. W. *et al.* Gene3D: structural assignments for the biologist and bioinformaticist alike. *Nucleic Acids Res.* **31**, 469–473 (2003).
119. Wilson, D. *et al.* SUPERFAMILY — sophisticated comparative genomics, data mining, visualization and phylogeny. *Nucleic Acids Res.* **37**, D380–D386 (2009).
120. Krishnamurthy, N., Brown, D., Kirshner, D. & Sjolander, K. PhyloFacts: an online structural phylogenomic encyclopedia for protein functional and structural classification. *Genome Biol.* **7**, R83 (2006).
121. Marchler-Bauer, A. *et al.* CDD: a conserved domain database for interactive domain family analysis. *Nucleic Acids Res.* **35**, D237–D240 (2007).
122. Heger, A. *et al.* PairsDB atlas of protein sequence space. *Nucleic Acids Res.* **36**, D276–D280 (2008).
123. Orengo, C. A., Stilltoe, I., Reeves, G. & Pearl, F. M. G. What can structural classifications reveal about protein evolution? *J. Struct. Biol.* **134**, 145–165 (2001).
124. Mizuguchi, K., Deane, C. M., Blundell, T. L., Johnson, M. S. & Overington, J. P. JOY: protein sequence-structure representation and analysis. *Bioinformatics* **14**, 617–623 (1998).
125. Dayhoff, M. O. & Eck, R. V. in *Atlas of Protein Sequence and Structure 1967–1968* 33–45 (National Biomedical Research Foundation, Silver Spring, Maryland, 1968).
126. Henikoff, S. & Henikoff, J. G. Amino acid substitution matrices from protein blocks. *Proc. Natl Acad. Sci. USA* **89**, 10915–10919 (1992).
127. Lee, S. & Blundell, T. L. Ulla: a program for calculating environment-specific amino acid substitution tables. *Bioinformatics* **25**, 1976–1977 (2009).
128. Letunic, I. & Bork, P. Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation. *Bioinformatics* **23**, 127–128 (2007).

Acknowledgements
C.L.W. was funded by a Biotechnology and Biological Sciences Research Council studentship. S.G. was supported by the BiO foundation. T.L.B. is funded by the Wellcome Trust.

DATABASES
PDB: http://www.rcsb.org/pdb/home/home.do
1gk8 | 1m2z | 1tme | 2fgf | 2ohxa | 3app | 5pal

FURTHER INFORMATION
The Blundell group's homepage: http://www-cryst.bioc.cam.ac.uk/
CATH: http://www.cathdb.info/
CDD: http://www.ncbi.nlm.nih.gov/Structure/cdd/cdd.shtml
CE: http://cl.sdsc.edu
Dali: http://ekhidna.biocenter.helsinki.fi/dali/start
ESSTs: http://www-cryst.bioc.cam.ac.uk/ESST
Gene3D: http://gene3d.biochem.ucl.ac.uk/Gene3D/
HOMSTRAD: http://www-cryst.bioc.cam.ac.uk/~homstrad
InterPro: http://www.ebi.ac.uk/interpro
JOY: http://www-cryst.bioc.cam.ac.uk/joy
MMDB: http://www.ncbi.nlm.nih.gov/Structure/MMDB/mmdb.shtml
PairsDB: http://pairsdb.csc.fi
PASS2: http://caps.ncbs.res.in/campass/pass2.html
Pfam: http://pfam.sanger.ac.uk
PhyloFacts: http://phylogenomics.berkeley.edu/phylofacts
Prints: http://www.bioinf.man.ac.uk/dbbrowser/PRINTS/
ProDom: http://prodom.prabi.fr/prodom/current/html/home.php
PROSITE: http://www.expasy.ch/prosite
PyMol: http://www.pymol.org
SCOP: http://scop.mrc-lmb.cam.ac.uk/scop
SMART: http://smart.embl-heidelberg.de
Superfamily: http://supfam.cs.bris.ac.uk/SUPERFAMILY
TIFRFAMs: http://www.jcvi.org/cms/research/projects/tigrfams/overview/
Toccata: http://www-cryst.bioc.cam.ac.uk/toccata/toccata.php
Ulla: http://www-cryst.bioc.cam.ac.uk/ulla

SUPPLEMENTARY INFORMATION
See online article: S1 (figure) | S2 (table) | S3 (table)
ALL LINKS ARE ACTIVE IN THE ONLINE PDF