Annu. Rev. Biophys. Biomol. Struct. 2002. 31:45–71
DOI: 10.1146/annurev.biophys.31.082901.134314

THE NATURAL HISTORY OF PROTEIN DOMAINS

Chris P. Ponting[1] and Robert R. Russell[2]
[1]*Department of Human Anatomy and Genetics, University of Oxford, MRC Functional Genetics Unit, South Parks Road, Oxford OX1 3QX, United Kingdom; e-mail: Chris.Ponting@Human-Anatomy.oxford.ac.uk*
[2]*EMBL, Meyerhofstrasse 1, Postfach 10 22 09, D69012 Heidelberg, Germany; e-mail: Russell@embl-heidelberg.de*

Key Words protein evolution, protein structure, sequence analysis, domain classification, function prediction

■ **Abstract** Genome sequencing and structural genomics projects are providing new insights into the evolutionary history of protein domains. As methods for sequence and structure comparison improve, more distantly related domains are shown to be homologous. Thus there is a need for domain families to be classified within a hierarchy similar to Linnaeus' *Systema Naturae*, the classification of species. With such a hierarchy in mind, we discuss the evolution of domains, their combination into proteins, and evidence as to the likely origin of protein domains. We also discuss when and how analysis of domains can be used to understand details of protein function. Unconventional features of domain evolution such as intragenomic competition, domain insertion, horizontal gene transfer, and convergent evolution are seen as analogs of organismal evolutionary events. These parallels illustrate how the concept of domains can be applied to provide insights into evolutionary biology.

CONTENTS

DOMAIN IDENTIFICATION

In recent decades the concepts of domains and domain families have risen to greater prominence within science. This has been due to an increasing realization that division of a protein's structure or sequence into domains often precedes reliable and accurate predictions of molecular function. A view of a multidomain protein's function as the sum of its constituent parts is obviously simplistic, as it ignores possible interdomain interactions and cooperative effects. Nevertheless this view does provide a first-approximation prediction that is amenable to investigation and subsequent refinement using experimental approaches.

Although domains now permeate descriptions of biology, definitions vary. In protein structure, a domain is often viewed as a compact, spatially distinct unit. In biochemistry, domains are frequently described as protein regions with assigned experimental functions, irrespective of their three-dimensional (3D) structures. In sequence comparison, domains are viewed from an evolutionary perspective and described as significantly sequence-similar homologs that are often present in different molecular contexts. These three views are compatible for the many cases where sequence-similar homologs adopt similar folds and possess comparable functions.

In this review, we consider how domains are identified and classified into sequence- or structure-based families related by common ancestry (homology). Function will not be used to define domains because most domain families contain representatives with different functions. Additionally, providing a standard definition of function is fraught with problems greater even than those that arise in defining domains (49). The classification of domains by homology raises many interesting questions concerning their evolution. In particular, how do the rates of change of gene structure, protein structure, sequence, and function vary? Can distinct domain families be related by common ancestry, even when their sequences, and even structures, differ radically? Finally, how did domains first arise and has "domain genesis" been occurring in relatively recent times?

Domains in Three-Dimensional Structures

The concept of a domain was first used to describe distinct regions of protein 3D structures. In the 1960s, the earliest enzyme structures of lysozyme (13) and ribonuclease (51) contained spatially distinct structural units, which were termed domains. Subsequent structures, such as pyruvate kinase (109), also showed a similar division into these units. Gradually, it emerged that such domains could recur either in different structural contexts [such as Rossmann folds in lactate and alcohol dehydrogenases (90)] or in multiple copies in the same polypeptide chain

(such as for trypsin, pepsin, and rhodanese). The more recent availability of large numbers of protein sequences and structures has resolved many initial domain-assignment ambiguities. In addition, many early domain assignments have been corrected. Ironically, one such case is the protein lysozyme, which was thought originally to contain two domains. Lysozyme is now assigned as a member of a single-domain family whose representatives can contain elaborations to the original single-domain ancestral fold (43).

Today, domains within protein structures are usually defined as spatially distinct structures that could conceivably fold and function in isolation (Figure 1). Many methods that exist assign protein domains based on 3D structure, most of which are based on geometric measures of compactness (e.g., 48, 101, 104, 110). However, the advent of structure comparison has enabled the important principle of recurrence to play a central role in domain definition. The observation of a similar structure within a different context is a powerful indicator of a protein domain, and this can be irrespective of whether a unit is spatially distinct. This important principle of recurrence has been implemented in at least one method (47), and of course it is central to methods of domain assignment based on sequence similarity, which are discussed in the sections that follow.

Domains in Protein Sequences

The arrangement of different domain types in protein sequences causes considerable difficulties in sequence analysis. This can be seen from consideration of the proteins represented in Figure 2. These each contain a pleckstrin homology (PH) domain and a src homology 3 (SH3) domain, but in both orientations (PH then SH3, and SH3 then PH). Src homology 2 (SH2) and SH3 domains co-occur in Vav and Rgs1, but in two combinations: SH3-SH2-SH3 and SH2-SH3-SH2. Comparison of any one of these sequences with databases generates a plethora of significant alignments, a few with proteins that possess the same arrangement of domains, many more with proteins with different domain arrangements, and some with multiple hits within the same sequence. Teasing out the evolutionary relationships among the different regions of multidomain protein sequences therefore requires careful analysis of each set of regions that possesses a distinct evolutionary history.

Such a set of regions represents a family of homologous domains. Although the detection of homologous domain sequences in databases requires sophisticated analysis tools such as BLAST (2, 96), these are freely available on the Internet and, more importantly, are easy to use. Detecting homologous sequences may employ pairwise methods or, more effectively, generalized profile (GP) or hidden Markov model (HMM) methods (41, 78). GPs and HMMs are "domain descriptors," meaning they can easily retrieve from databases all domain sequences that make up their corresponding multiple alignments, as well as other homologs that might not have been previously thought to be members of the domain family.

Not all sequence families represent domains. Some sequence- and structure-similar entities are too small, or are lacking in secondary structures, to represent

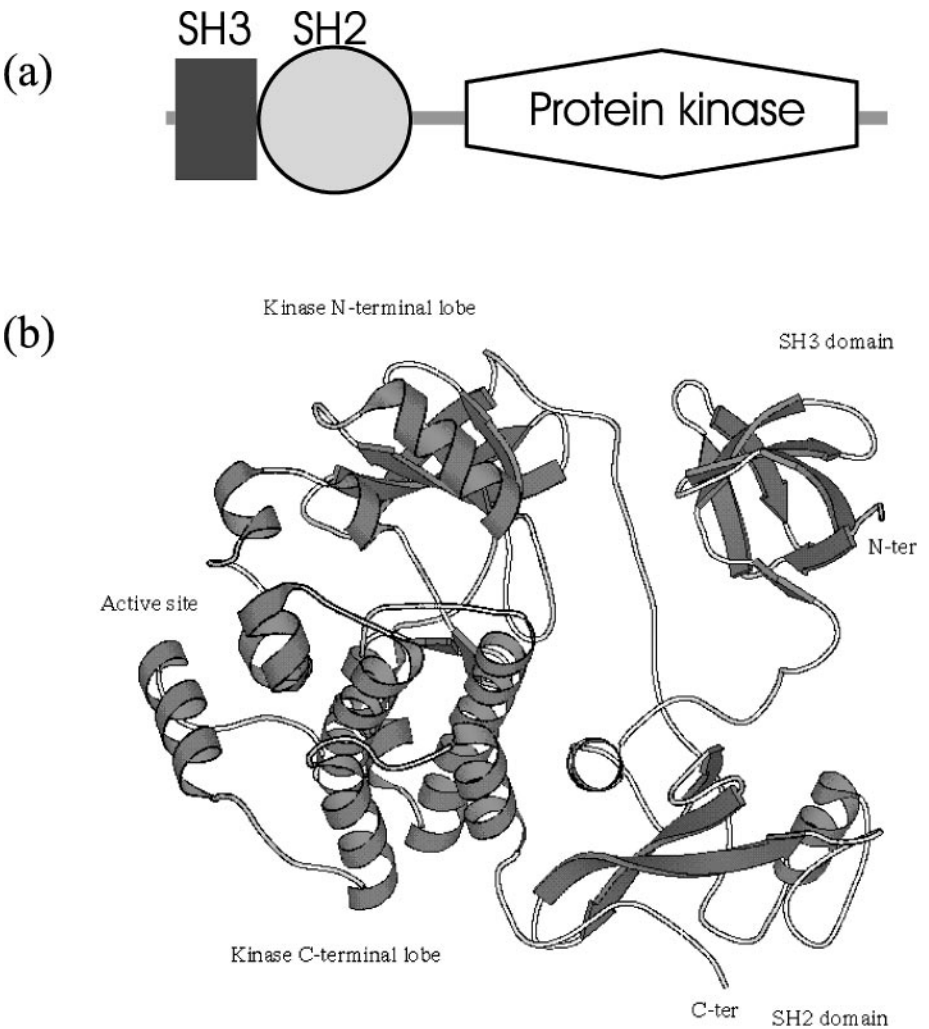


Figure 1 (*a*) Domain architecture and (*b*) Molscript (54) representations of a fragment of the structure of hematopoetic cell kinase (hck; PDB code 1qcf), containing protein kinase and src homology 2 and 3 (SH2 and SH3) domains.

domains. Examples of such motifs are DNA-binding AT-hooks (5) and short sequences that specifically target proteins to subcellular localizations (71). Some structural domains are entirely composed of repetitive structures of varying numbers (3). These repeat families can be classified into those that form linear rods (e.g., in spectrin) or superhelices (e.g., HEAT repeats) or closed structures (e.g., β-propellers or β-trefoils).

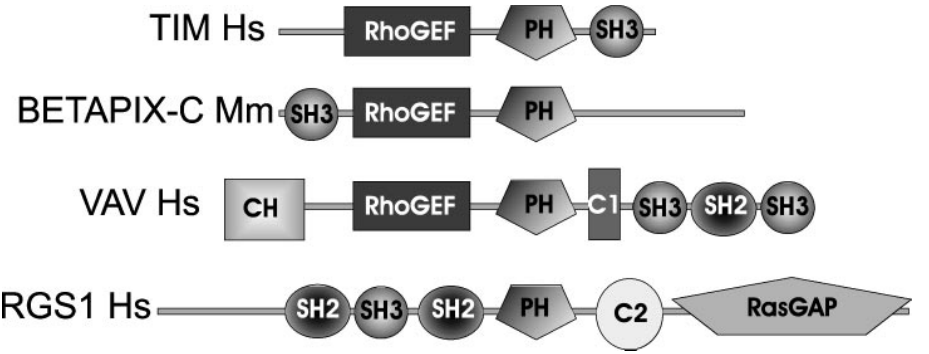


Figure 2 Representations of the domain architectures of human p60 tim (transforming immortalized mammary oncogene), mouse βPix-c, human Vav oncogene, and human RGS1 (regulator of G-protein signalling 1). Domain abbreviations: C1, protein kinase C conserved region 1; C2, protein kinase C conserved region 2; CH, calponin homology; PH, pleckstrin homology; RasGAP, GTPase activator protein specific for Ras-like small GTPases; RhoGEF, guanine nucleotide exchange factor specific for Rho-type small GTPases; SH2, src homology 2; SH3, src homology 3. Species abbreviations: Hs, *Homo sapiens*; Mm, *Mus musculus*.

Libraries of Domain Sequences

Libraries of domain, repeat, and motif alignments, and their associated GPs or HMMs, are available to automatically assign domains in protein sequences. These have become invaluable tools in sequence analysis, particularly in domain-based analyses of completely sequenced eukaryotic genomes (e.g., 20, 56, 115). They represent the current knowledge of domain families (thus avoiding much tedious repetition of previous analyses), and they allow results to be provided quickly and automatically.

These domain libraries have their own specialities. Of the HMM libraries, SMART [http://smart.embl-heidelberg.de/ (97)] and TIGRFAMs [http://www.tigr.org/TIGRFAMs/]; (39) focus on in-depth eukaryotic and prokaryotic families, respectively, and Pfam [http://www.sanger.ac.uk/Pfam/ (8)] is intent on providing comprehensive domain annotation of all proteins. Prosite provides a search of GPs, centered mostly on eukaryotic domains [http://www.isrec.isb-sib.ch/software/PFSCAN_form.html (42)]. To obtain all possible predictions, a user should visit each of these sites in turn. However, the ready availability of these libraries has spawned several meta-sites, such as InterPro [http://www.ebi.ac.uk/interpro/ (4)], CDD (http://web.ncbi.nlm.nih.gov/Structure/cdd/cdd.shtml), and Panal [http://mgd.ahc.umn.edu/panal/ (102)], that enable searches of several libraries to be performed simultaneously.

By assigning domains to homologous groups, these libraries provide a measure of objectivity. Where there is subjectivity is in the definition of what constitutes a domain family. In a model of evolution where domains evolve from antecedent

domains rather than de novo, decisions need to be taken regarding whether a sequence belongs to one subfamily rather than a second related subfamily, and regarding when a sequence is best considered to be within a larger family encompassing both such subfamilies. For example, SMART primarily predicts mouse Rho-A to be a member of the Rho-type subfamily of small GTPases (Figure 3). However, as Rho-type small GTPases represent a subfamily within the larger family of Ras-like GTPases, Rho-A is also viewed by SMART as being more distantly related (i.e., with less significant statistics) to Rab, Ras, Arf, and Ran small GTPases. On the other hand, Pfam assigns Rho-A to the large Ras family defined to encompass all of the Ras, Rab, Rac, Ral, Ran, Rap, and Ypt1 subfamilies of small GTPases.

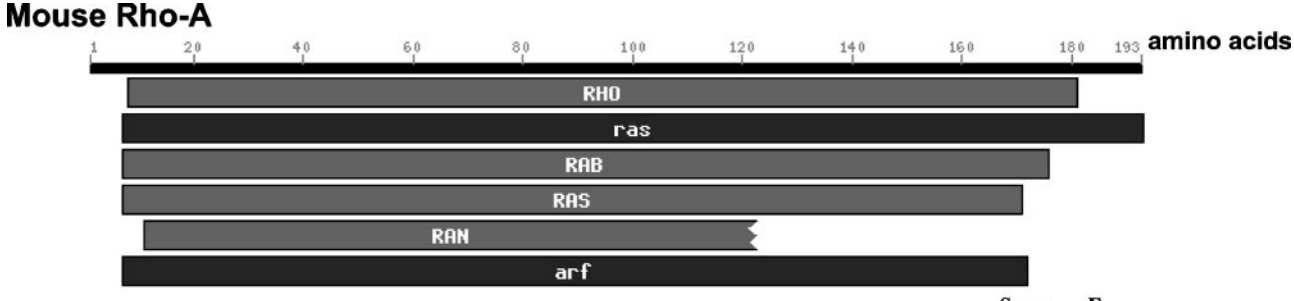
Thus a hierarchy is required to represent each of these evolutionary relationships: that Rho-A is a Rho-type small GTPase of the Ras family, which, in addition, is a member of a larger superfamily of GTPases and ATPases including dynein, myosin, and elongation factor Tu. The establishment of such hierarchies within domain libraries would provide significant added value to these resources. However, at present this development remains in its infancy.

Structure and Sequence Conservation

Domain recurrences among 3D structures consistently reveal that protein structure is more conserved than sequence. There are many examples of domains adopting highly similar 3D structures despite no apparent similarity in sequence. For many of these examples, proteins have diverged beyond the limits of sequence similarity detection methods but have nevertheless retained a common structure and similar function. For example, adenylate cyclase and DNA polymerase contain a similar domain that was recognized by 3D structure comparison (7). Despite a lack of obvious sequence similarity, both domains contain the active sites of these enzymes, and conserved residues are involved in catalyzing a similar reaction.

It is now usually accepted that proteins sharing a common fold and showing signs of a common ancestor reside in the same superfamily. Murzin and coworkers have been instrumental in making use of unusual features to classify proteins as being remote homologs. The most obvious indicators of common ancestry are features relating to function, such as catalytic or binding sites (e.g., 7, 15, 46, 68). However, other unusual features can also be used, such as left-handed β-α-β motifs (67) or other unusual topological connections unlikely to arise multiple times in the same structural position.

Figure 4 shows an example of such a feature. We argue that four superfamilies within the "swiveling" $\beta\beta\alpha$ domain (60) are homologous owing to the presence of an unusual loop connecting two β-strands at the edge of the core secondary structure elements that make up this fold. For two other superfamilies this feature is absent. Consequently, this unusual feature unites some, but not all, proteins adopting this fold.



		Score (bits)	E value
gnl|Smart|smart00174	RHO, Rho (Ras homology) subfamily of Ras-like small GTPases; M...	322	9e-90
gnl|Pfam|pfam00071	ras, Ras family	216	1e-57
gnl|Smart|smart00175	RAB, Rab subfamily of small GTPases; Rab GTPases are implicate...	139	2e-34
gnl|Smart|smart00173	RAS, Ras subfamily of RAS small GTPases; Similar in fold and f...	116	1e-27
gnl|Smart|smart00176	RAN, Ran (Ras-related nuclear proteins) /TC4 subfamily of smal...	65.9	2e-12
gnl|Pfam|pfam00025	arf, ADP-ribosylation factor family	44.7	5e-06

Figure 3 Web-based output from the comparison of the mouse Rho-A sequence (193 amino acids long) against the Pfam and SMART libraries of domains using the CDD (http://web.ncbi.nlm.nih.gov/Structure/cdd/cdd.shtml) system. Rho-A is seen to be most similar to the Rho (Ras homology) subfamily of Ras-like small GTPases as comparison with this domain alignment yields the lowest Expect (*E*-) value of 9×10^{-90}. If an alignment results in a score *x*, then its associated *E*-value represents the number of sequences expected to be aligned against the query with scores *x* or greater in this search simply by chance.

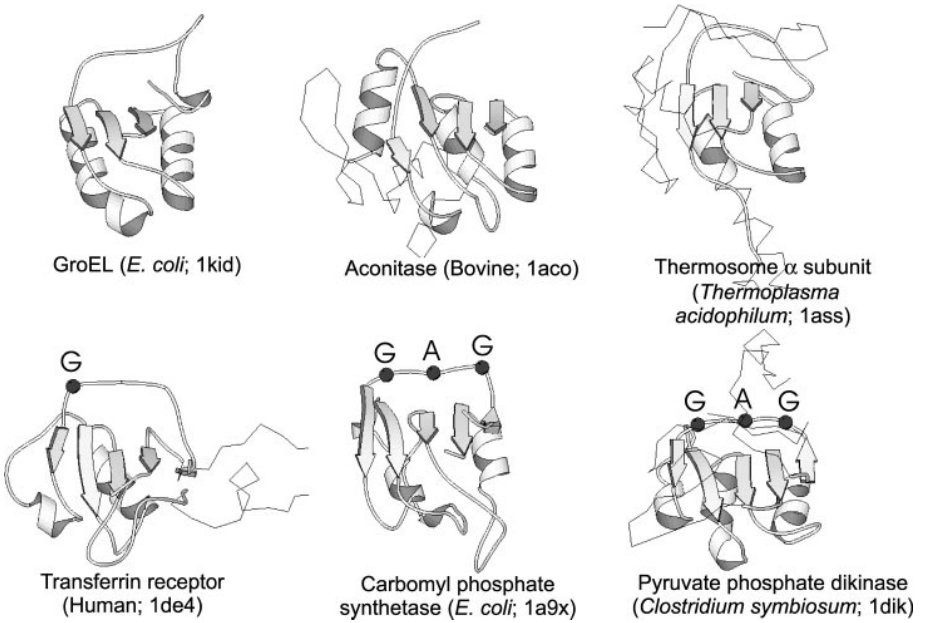


Figure 4 Molscript (54) figures showing the relationship between proteins from six superfamilies that adopt the "swiveling" $\beta\beta\alpha$ domain. The unusual loop connecting two edge β-strands in carbomyl phosphate synthetase, pyruvate phosphate dikinase, transferrin receptor, and GroEL suggests that they may be homologous. The core structure common to most proteins is shown as arrows (β-strands), ribbons (α-helices), or coil, with additional regions shown in C_α trace. Spheres show the location of conserved glycine (G) or alanine (A) residues found in some members containing an unusual loop as discussed in the text.

Attempts to automate superfamily assignment have failed to detect all such relationships but have nevertheless gone some way toward faster recognition of homology. Approaches have used sequence similarity (see below), the definition of a homologous core structure (64), or a combination of approaches, including sequence bridges or functional annotations from sequence databases (45, 62).

In recent years methods have been developed that use structure and sequence together to assess whether pairs of sequence-dissimilar domains that adopt the same fold are homologs. Russell et al. (94) found that when comparing only the structurally equivalent positions of such domain pairs, sequence identities of 12% or higher were likely to indicate homology. Murzin (66) devised a probability measure for assessing the significance of structurally equivalent and identical residues and used this to argue for a common evolutionary origin of cystatins and monelins. More recently, this measure was employed for determining a common ancestor for β-trefoil proteins, such as interleukin-1 (IL-1), fibroblast growth factors, and actin-binding proteins (87).

Classification of Three-Dimensional Structures

Over the past eight years, several schemes for protein structure classification have developed and matured and are now widely used for studies of protein structure, function, and evolution. It is beyond the scope of this review to discuss them all in detail, but salient points of three main schemes are discussed here. All classifications attempt to build a hierarchy of protein structural domains and highlight instances when protein structures show evidence of a common ancestry despite low sequence similarity.

Murzin et al. maintain and manually curate the structural classification of proteins (SCOP) database [http://scop.mrc-lmb.cam.ac.uk/scop/index.html (60, 69)]. The authors often point out that SCOP is an evolutionary classification, the main focus being to place proteins in the correct evolutionary framework, based on conserved features discussed in the previous section. Manual curation of SCOP means that sometimes it is not as up-to-date as the Protein Data Bank (PDB) and that some protein folds have been the subject of more attention than others. Nevertheless, the limits of its construction should not be overstated because it remains a key resource applicable to many purposes.

Orengo et al. maintain the CATH database [http://www.biochem.ucl.ac.uk/bsm/cath_new/index.html (76)]. This is constructed using a mixture of manual and automated methods and contains excellent supplementary information and cross-referencing for all structures. Information concerning active sites and homologous sequences is also readily available.

Holm & Sander offer the fully automated fold classification based on structure-structure alignment of proteins (FSSP) database [http://www.ebi.ac.uk/dali/fssp (44)], which is as up-to-date as the PDB. Automated procedures for domain assignment and protein structure comparison group proteins into clusters sharing similar folds. The authors have also provided a means for assigning homology despite low sequence similarity (45).

Despite differences in philosophy and design, the three classifications often agree (38). However, as is the case for domain assignment from sequence, the methods and details are sufficiently different to warrant inspection of their results in cases when one is attempting to place a structure in the correct structural, functional, or evolutionary context. This is particularly true in instances when either domain assignment or structural similarity is ambiguous.

EVOLUTION OF DOMAINS

Sequence-based analyses have demonstrated that some domains have ancient origins because they are widespread in each of the three forms of cellular life, archaea, bacteria, and eukarya, whose common ancestor existed over three billion years ago. The persistence of such domains implies that they are either hyperadaptable, suited to many beneficial functional niches, or that they are essential to fundamental

cellular processes. Many enzymatic domains of central metabolism [e.g., $(\beta/\alpha)_8$ TIM barrels, flavoproteins, and Rossmann-like fold proteins] (32, 119) appear to owe their heritage to ancestors that preceded the last common ancestor of archaea, bacteria, and eukarya (30).

Ancient domains are not all enzymatic. PSD-95, Dlg, Zo1 (PDZ) domains, for example, occur in bacterial periplasmic proteins and archaeal tricorn-like proteases, as well as in numerous eukaryotic signaling proteins (82, 85). The binding of nonpolar protein C termini to PDZ domains is conserved among prokaryotic and eukaryotic proteins (10). Thus the molecular function of this domain family has been retained over several billion years. Other nonenzymatic domains are also likely to have been present in the last common ancestor of cellular life, including cystathionine β-synthase, plant pathogenesis-related-1, and von Willebrand factor A domains (84). Whether these ancient domains also retain conserved functions remains unknown.

Other domain families appear to be eukaryotic inventions because homologs cannot be detected in known prokaryotic sequences. Enzymatic and nonenzymatic domains of ubiquitin-mediated proteolysis, actin-binding cytoskeletal domains, and chromatin-associated domains, among many others, are represented only in eukaryotic proteomes (88). Other domain families, such as DEATH, DED, CARD, and PYRIN, are specific to eukarya that exhibit apoptosis and consequently are thought to be metazoan inventions. All four domains possess similar dimerization roles and similar tertiary structures (118). Thus they have arisen from a common ancestor, early in metazoan history, and have since diverged in sequence. This demonstrates the limitation inherent in defining domain families using sequence information alone.

Many extracellular domain families (e.g., apple, CCP, C-lectin, furin, fibronectin type-1 and -2, Gla, and kringle) are represented in metazoan proteins when they are absent elsewhere. The evolution of multicellularity in animals occurred concurrently with the proliferation of extracellular domain types that would have been required for cell-cell communication roles. Many of these metazoan extracellular domains are absent from known plant sequences, which suggests that multicellularity in plants and animals evolved independently (6, 23).

Domain Origin and Antecedent Domain Segments

Cytokines, such as IL-1α, -2, -4, -7, -9, -10, and -13, and the immunoglobulins required for the acquired immune response appear to have arisen in more recent times, since the emergence of chordates (56). All such discussions of domain emergence beg a fundamental question: From what did these domains evolve? Were they products of preexisting domain families that became extinct or else considerably diverged in sequence? Did such domains originate from noncoding DNA predecessors? The fundamental difference between these alternatives is that the former assumes a continuous evolution and vertical descent of domains, whereas the latter

describes a discontinuity of domain evolution, a "domain parthenogenesis," where novel domains are generated ab initio.

A study of IL-1α, a β-trefoil-fold domain (87), using the structure- and sequence-dependent similarity method of Murzin (66) described previously, recently showed that this cytokine arose from a protein precursor homologous to invertebrate fibroblast growth factors, and slime mold and fungal actin-binding proteins. This illustrates a conclusion common to many findings (e.g., 9, 98): The more the sensitivity of sequence and structure database searching algorithms improves, the more it is realized that previously isolated domain families are instead part of larger, evolutionarily related superfamilies.

Consequently, as fast as new domain families are being found, old family ties are being discovered. There might appear to be a fundamental limit to this merging procedure because domain families with different folds are not usually considered as potential homologs. However, there is evidence that not only domains, but also their folds, might have radically changed. If so, then this suggests that most modern folds might have arisen from, at most, a few ancestors.

Regardless of how many times domains with different folds have arisen, there is little explanation of what were the precursors of the earliest domains. Recently, Lupas et al. (61) considered a variety of phenomena and suggested that domains may be descended from conglomerates of short polypeptide segments that together were capable of folding and conveying a beneficial function. Evidence for the past importance of short polypeptides was twofold and was gathered from modern proteins that contain either homologous repeats or short sequence-similar motifs embedded in nonhomologous structures.

First, it was argued that proteins containing multiple copies of a homologous repeat, by definition, must be derived from a short polypeptide ancestor that possessed only a single repeat. In order to possess a biological function, it is argued, this ancestral repeat must have oligomerized in order to adopt a structure comparable to that of modern homologs. For example, a study of IL-1α homologs (87) revealed that the three pseudosymmetrical regions of the β-trefoil fold are homologous, having arisen from a common ancestor. The implication is that this common ancestor once adopted a β-trefoil fold by forming a trimer of identical subunits and that its descendants arose by intrachain subunit duplication. A similar argument may be made for all internally repetitive domains, such as β-propellers (65) and triple β-spirals (114).

Second, short highly sequence-similar motifs, such as Asp-box and helix-hairpin-helix (HhH) motifs (22, 28), have been identified in nonhomologous structures. One model for the evolution and spread of these unusual short sequence and structure motifs is that they represent ancient conserved domain cores that have persisted owing to their relatively high importance for function and structure while their surrounding structures have been subjected to greater alteration. Another model is that they represent short gene segments that have been successfully duplicated and incorporated into different nonhomologous contexts.

Whether such motifs are ancient or more modern, their existence indicates that domains might be divisible and have arisen by recombination of smaller sequences.

The past occurrence of short polypeptides, which are seen today as internal repeats and structure-integrated motifs, argues that complex single-domain structures might have arisen by the fusion of simpler substructures, much in the same manner as complex multidomain proteins are thought to have arisen by domain shuffling (61). The ancient predomain world would thus have contained much shorter proteins [antecedent domain segments (ADSs)] that would have acted in concert to produce single-domain homo- or hetero-multimers. This idea agrees with preconceptions as to the nature of the early protein world where, if proteins were as long as often they are today, higher error rates in transcription and translation would have been likely to produce misfolded variants.

A degree of support for this ADS hypothesis comes from recent improvements in protein tertiary structure prediction. Bystroff et al. and Simons et al. (16, 17, 103) have made use of a library, containing local regions of structure similarity (I-sites) common to different protein folds, to obtain outstanding prediction accuracies. These I-sites might correspond to ancient ADSs because their prediction success might have arisen from an identification of regions of genuine ancient homology among otherwise different 3D structures.

Correspondence Between Exons and Three-Dimensional Structure

Since exons were first identified it has been proposed that they might correspond to units of protein 3D structure or function (12, 29, 33). An underlying assumption to this proposition is that the combination of exons was once key to the construction of protein structures seen today (29). Many studies have been performed during the past two decades searching for a link between intron positions and units of 3D structure or function (e.g., 25, 107, 113).

The role of introns in evolution has been the subject of much controversy. In one camp are adherents of the introns-early theory (e.g., 25, 34). They argue that introns were present in the progenitor of all living organisms and were subsequently lost in bacteria and archaea. For this theory to be true, one expects, at least for ancient proteins, that intron positions will lie at key junctions in protein structures, suggesting their original assembly by exon-shuffling, the evidence for which is now missing in bacteria and archaea.

In the opposite camp are those who believe that the absence of introns from bacteria and archaea reflects the fact that introns are eukaryotic inventions (19, 77, 81, 99). A perceived failure to correlate exons with units of structure or function, the apparent recent origin of exons, and recent evidence highlighting key roles for exons during various stages of eukaryotic evolution (see below) have led many to argue that introns have arrived only recently in eukaryotic evolution (the introns-late hypothesis).

There have been many studies during the past two decades attempting to resolve this debate. In summary, much of the community agrees that the introns-early theory is untenable (e.g., 107), although de Souza and coworkers continue to argue that there is evidence that certain key exons likely represent ancient building blocks for many ubiquitous proteins (25). The sudden availability of thousands of genomic sequences for proteins is likely to resolve this debate soon.

Despite controversy as to the role of introns in the origin of ancient proteins, there is a growing body of convincing evidence for the role of exon-shuffling in the assembly of recently evolved proteins in eukaryotes. For example, Patthy (79–81) has provided convincing evidence for a role in exon-shuffling during the "big bang" of metazoan radiation. This period saw the emergence of the first multicellular animals, an event that is correlated with the first appearance of many new extracellular signaling proteins. These proteins, which are needed to mediate communication between cells in multicellular organisms, are often complex combinations of many different domains. Many show a good correlation in the location of domain boundaries and the location of introns, particularly those of phase 1. (The phase of an intron is defined as 0, 1, or 2, where the number refers to the location of the intron relative to the nucleotide triplet; e.g., 0 implies that the intron does not interrupt a codon.) For this reason, Patthy argues that exon-shuffling played a key role in the construction of multidomain extracellular signaling proteins. This is in contrast to proteins found inside the cell (such as kinases, SH2, SH3, and PH domains) where there is scant evidence for correlation. He thus argues in addition that such intracellular signaling molecules probably first arose prior to the emergence of exon-shuffling as a major evolutionary force.

Recently, Betts et al. (11) investigated the degree to which intron positions are conserved within protein domains rather than within multidomain proteins. They found a surprising number of families where intron positions were conserved despite little or no sequence similarity. Here similarity between proteins was only inferred following 3D structure comparison. Most intriguingly the proteins that showed conservation of intron positions were often in the same functional class (such as immune system proteins) or restricted to particular evolutionary lineages (such as nematodes). This suggests that such intron conservation is a relic of comparatively recent divergence, followed by rapid evolution owing to key events in evolutionary history, such as the developments of the immune system in vertebrates or the chemical response system in nematodes. It also raises the possibility that unequal crossing-over (essentially exon-shuffling) might have created domain hybrids, thus increasing the rate at which these domain families could diverge.

Fold Changes During Domain Evolution

It has long been clear that proteins adopting similar folds often differ from each other outside of a conserved core. Homologs can have additional segments at their N- or C-terminal ends, and loop regions often differ, particularly when sequence

similarity is low and functions are different. Sometimes loop regions are the site of domain insertions, where domain duplications have led to one domain being copied into the middle of another (see below). Another simple but drastic evolutionary event is circular permutation, which presumably occurs by gene duplication, fusion, and partial deletion (93) and can lead to substantial changes to the topology of a protein fold. Conceptually it can be considered to involve a fusion of the N and C termini and a cleavage at a different location to create new termini. It is thus only strictly permitted in proteins where N and C termini are close to each other in space.

It is also possible for regions of proteins to change dramatically in structure with conventional mutation and insertion events. The landmark work of Grishin (37) has demonstrated that the current population of protein structures contains numerous instances where seemingly different protein structures can be argued to have descended from a common ancestor. It is possible, for example, for regions of α-helix to change over time to become short β-sheets. It has even been proposed that it is theoretically possible for an evolutionary transition to occur between all-α and all-β protein structures by a series of events already seen in the current set of protein structures.

Convergent Evolution

Although there is growing evidence that an increasing number of apparently different structures may share a common ancestor, it is also clear that nature has reinvented, by convergent evolution, similar local structures multiple times. Probably the best-known example is that of the Ser/His/Asp catalytic triad (27), which is found in at least five different protein folds that cannot easily be considered to be homologous by any of the evolutionary events discussed above. It is likely that nature has been limited by the 20–amino acid alphabet in the choice of residues that can perform similar functions.

Methods detecting recurrent 3D side chain patterns have found numerous instances of localized structure convergence (e.g., 92, 117). Convergence of function has also been observed within homologous protein families. In these cases nature has invented a substrate specificity more than once within the same homologous family (e.g., 24, 120). One of the more fascinating cases of structure convergence concerns thermolysin and mitochondrial processing peptidase (63). These proteins show striking similarities not only in their active site residues but also in their structures. The arrangement and packing of the core secondary structure elements is the same, yet the connectivity (i.e., the order of the elements along the polypeptide chains) is completely different, often in the reverse main chain orientation, thus making their descent from a common ancestor extremely unlikely.

It remains an open question as to whether whole-protein folds, rather than localized structures and functional sites, have arisen multiple times during evolution. Nature could have stumbled upon simple folds, such as four α-helical bundles, multiple times, but for more complicated structures this is not so easy

to argue convincingly. By contrast, several folds [such as β/α-(TIM)-barrels and β-trefoils], which were previously thought to be examples of convergence, are increasingly being revealed as homologs (21, 87) due to enhanced sensitivity in sequence comparison techniques.

Domains and Protein Evolution

Duplication of genes within a genome has been a major evolutionary process in the acquisition of novel function (74). The initial functional redundancy after duplication reduces mutational constraints for one or both copies and gives rise to an increased likelihood of functional differentiation. Gene duplicates that have persisted to the present day are known as paralogs; this contrasts with orthologs, which are genes that arose by speciation rather than by intragenome duplication. Partial gene duplication has also been a prominent mechanism for function diversification. Frequently, this has generated multiple tandem repeats or else the occurrence of homologous domains in different domain architectures. A third prominent evolutionary mechanism is deletion, either deletion of genes or portions of genes.

Domain duplication and acquisition lead to variations in the tandem arrangement of domains along a sequence. However, duplicated domains are not always inserted into genes in a manner that results in tandem domains. In relatively rare instances, duplication can give rise to the insertion of a domain within another domain. For such cases whose tertiary structures are known, the fold and structural integrity of the two inserted and parent domains remain intact. The polypeptide backbone of the parent domain is interrupted by an excursion from an external loop to form the inserted domain before returning to complete the parent domain fold (91). This is made possible, it is thought, by the frequent spatial proximity of N- and C-terminal ends of protein domains. Detecting cases of domain insertion using sequence analysis is problematic (93). However, among documented cases, several involve PH domains. PH domains can have embedded C1 and PDZ domains (in Rho-associated kinase α and syntrophins, respectively) or else be inserted into a band 4.1 domain or another PH domain (in Mig2 and myosin X, respectively) (93).

The importance of duplication and deletion events in past evolution can be inferred from the repetitive and piece-meal nature of modern proteins. This is illustrated here by a family of proteins typified by mouse Ky (14) and leech hillarin (50) (Figure 5). The members of this protein family are closely related to each other in sequence, yet their domain architectures differ in key respects. Leech hillarin is duplicated with respect to mouse Ky, and *Caenorhabditis elegans* and *Saccharomyces cerevisiae* Ky homologs have independently acquired LIM and SH3 domains, respectively. The effect of these duplication and domain acquisition events on molecular function remains unknown. Although the C-terminal regions of these genes are likely to be orthologous, having arisen from a common ancestor by speciation events, the lineage-specific evolution of domain architectures implies that their molecular functions have diverged significantly. This matter is made more

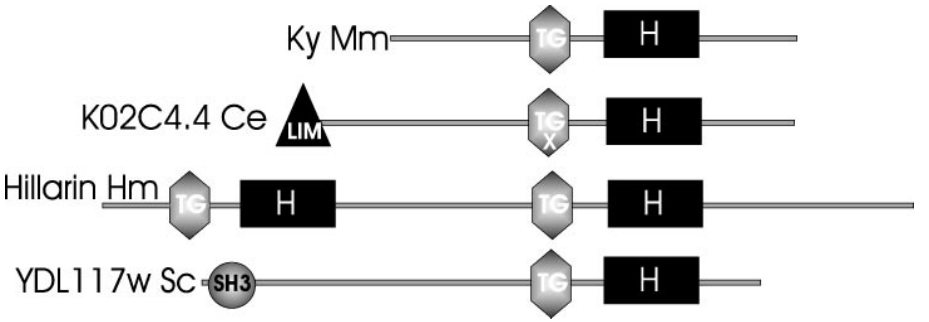


Figure 5 Representations of the domain architectures of *Mus musculus* (Mm) Ky homologues from *Caenorhabditis elegans* (Ce), *Hirudo medicinalis* (Hm), and *Saccharomyces cerevisiae* (Sc). Conserved transglutaminase (TG) homology domains and H domains (50) are shown, together with src homology 3 (SH3) and Lin-11, Isl-1, Mec-3 (LIM) domains.

complex by the realization (14) that the *C. elegans* Ky homolog alone among these proteins has substitutions within the putative active site of its transglutaminase homology domain, implying a further divergence of their functions.

Differences in domain architectures among genes that have clearly arisen, at least in part, from a common ancestor raise the question of whether these genes are orthologous. If, as is likely, the Ky ancestor contained only transglutaminase and H domains, then only the regions of leech, mouse, *C. elegans*, and yeast Ky homologs that contain transglutaminase and H domains are orthologous (Figure 5). The remaining regions that were absent from the common gene ancestor, a SH3 domain in yeast and a LIM domain in *C. elegans*, appear to be later additions and consequently are not orthologous or even homologous. Similar to the conventional use of the concept of homology, it is suggested that descriptions of orthology are most appropriately applied to domains, rather than proteins, except when proteins contain identical domain architectures. Although this may appear at first to be a question of semantics, the identification of orthologs is crucial for the comparison of function across species (52). Furthermore, the comparison of function across species is central to the issue of whether science is to take full advantage of the newly sequenced genomes.

THE ROLE OF DOMAINS IN PREDICTING FUNCTION

Domains and Organismal Function

Libraries of domain families, in particular Pfam and SMART, have recently been used to great effect to provide insights into eukaryotic evolution (56, 115). In one of these studies (56), 94 domain families were identified to have arisen in the chordate lineage and reflect the emergence of novel physiological systems. Of

these, 23 function in the defense and immunity systems (e.g., ILs), 18 are plasma factors (e.g., uteroglobin), 17 are found in the peripheral nervous system (e.g., ependymin), and 14 are involved in bone and cartilage formation (e.g., calcitonin), lactation (e.g., caseins), or vascular (e.g., endothelin) or dietary (e.g., glucagon) homeostasis. Clearly these families are markers of the chordate lineage and reflect the participation of emergent domain types in the evolution of the immune, nervous, skeletal, mammary, vascular, and digestive systems in chordates in general, and vertebrates in particular.

These 94 domains, however, are a small minority of all families represented in vertebrates. The great majority of domain families (over 90%) originated prior to the appearance of the first chordates (56). Yet, although most vertebrate domains are ancient in origin, the proteins in which they appear often are not, having been constructed by combining domains into an abundance of different architectures: The number of such architectures in vertebrates exceeds that in invertebrates by about 80% (56). The recent availability of the complete genome sequences for *Homo sapiens*, *Drosophila*, *C. elegans*, *Arabidopsis*, and *S. cerevisiae*, therefore, has demonstrated the significant contribution made to the evolution of organismal function by the variation of domain combinations.

Thus far, the propagation of genes and domains has been described only as a process of transferral by vertical descent. However, considerable evidence has been amassed that among prokaryotes, and bacteria in particular, genes have been passed between genomes (31, 72). This horizontal gene transfer among genomes stymies our attempts to understand the relatedness of anciently diverged organisms but may provide insights into the evolutionary relationships between coexisting organisms, such as pathogen and host or symbiont pairs (55, 73). Clearly, a gain by horizontal gene transfer to the germ line of a domain that confers advantages to the acquisitive organism enhances the perpetuation of its genome.

At first, it appeared that the human genome might have acquired domains and genes via horizontal gene transfer from bacterial sources (56); this is now considered less likely (106). Horizontal gene transfer from vertebrates into bacteria, however, certainly has occurred (84). Bacteria are likely to have benefited from acquisition of domains that, in vertebrates, function in cell-cell signaling (integrin, cadherin, and fasciclin domains), intracellular signaling (Sec7 domain and WD40 and ankyrin repeats), and chromatin remodeling [Su(var)3-9, enhancer-of-zeste, trithorase (SET) domain]. Each of these acquired domains is likely to have been fixed in its recipient bacterial genome as a result of a direct benefit conferred to the organism. One approach to enhancing our understanding of pathogenicity in vertebrates therefore is to identify pathogen genes gained from host organisms by horizontal gene transfer. One such example is the apparent acquisition by the intracellular human pathogen *Chlamydia* of a perforin-homologous gene from a vertebrate source (83).

The principle of natural selection is applicable to many different biological objects (36), including clades, species, individuals within species, and single genes within individuals. Could one such biological object be a domain family? We

believe that there is evidence for intragenomic competition among domain families for particular functional niches.

Consider the example of protein kinases. These may be divided into two major families: (*a*) histidine kinases (HisKs) that phosphorylate predominantly on histidine and (*b*) kinases that phosphorylate on serine or threonine or tyrosine (STYKs). Each of these families is involved in intracellular signaling, and each contains an ATPase active site that appears to have arisen from a common ancestor (53). However, HisKs are almost exclusively limited to bacteria and archaea, with a few present in fungi and plants, and STYKs are almost exclusively limited to eukarya, with the exception of antibiotic kinases and bacterial kinases acquired from eukarya via horizontal gene transfer (58). For example, *Escherichia coli* has 30 HisKs and no STYKs, whereas *H. sapiens* has over 550 STYKs and only 4 extremely distant HisK homologs (pyruvate dehydrogenase kinases). Phylogenetic analysis evidence suggests that eukaryotic STYKs are relatively recent inventions, arising since the last common ancestor with bacteria and archaea, and that HisKs are ancient, being present in the last common ancestor of the three forms of cellular life. This suggests that, in animals where STYKs are abundant, the success of this family of kinases in colonizing intracellular signaling niches has led to the virtual extinction of HisKs. This effect may be simply due to an enhanced stability of phospho-Ser/Thr/Tyr over phospho-His, which is required for larger cells, where signals need to be transduced over greater distances.

A more contemporaneous example of interfamily rivalry might involve the PH domain family, which is specific to eukarya. The archetypal PH domain family is known to be multifunctional, as it associates with membranes, inositolphosphates, and proteins (100). In eukaryotes, five major intracellular signaling modes exist that are mediated by the binding of phosphoserine or threonine, phosphotyrosine, polyproline, phospholipids, or small GTPases. Of these, the latter four modes are performed by domain families (PTB, WH1, PH, and RanBD domains, respectively) that adopt the PH domain fold. As these domain families occur only in eukaryotes and the PH domain fold has not yet been found outside of eukarya, there is reason to believe that each of the PTB, WH1, PH, and RanBD domain families arose from a PH domain fold common ancestor that arose early in eukaryotic history. The four signaling modes are also performed by other domain families (e.g., SH2, SH3, C2, and RA domains, respectively) that do not adopt the PH domain fold and also arose early in eukaryotic evolution. Thus it would appear that the PH domain fold has successfully colonized each of the four different signaling modes in direct competition with other domain folds. The basis for this hyperadaptability of the PH domain fold remains unclear.

Domain Families and Function

Knowledge that a protein adopts a particular protein fold is often insufficient to infer details of function (75). As discussed in previous sections, proteins with no apparent similarity in function can adopt similar 3D structures. Sometimes structure similarity does imply an ancient evolutionary relationship and often a

similar function, but for other structure similarities the situation is not so clear. Even in the absence of clear indicators of common ancestry or function, it is still possible to predict details of function if the protein adopts one of the folds known to show a preference for binding ligands in a particular location. Russell et al. (95) identified nine folds that showed a statistically significant tendency to bind ligands in a common location. This they termed a supersite because it occurs by definition in a protein adopting a superfold (75). β/α-barrels, doubly wound Rossmann-like folds, β-propellers, four α-helical bundles, ferredoxin-like folds, and others have a preferred location for binding to ligands that could well be dictated by principles of protein structure rather than by a common ancestry.

Although less useful than identifying a clear case of homology, knowledge that a protein adopts a fold containing a supersite can assist experimental design. The bacterial periplasmic protein TolB provides an example. TolB was predicted to contain a β-propeller domain, and the knowledge that β-propellers predominantly bind ligands in a common location led to a prediction of the TolB-binding site (86). This prediction corresponds well with several amino acids involved in suppressor mutations of *pal* A88V (89).

The completion of genomic sequencing projects means that many proteins of known sequence are of unknown function. In addition, high-throughput structural biology projects ("structural genomics") are producing 3D structures of proteins, whose functions remain unknown. Because comparison of sequence and structure is key to predicting function for new proteins, understanding the degree to which functional information can be transferred from one protein to associated homologs remains a major issue. An important question that has been addressed recently is how functional information can be transferred from one domain to its homologs. Todd et al. (112) investigated functional similarity for 31 diverse superfamilies in the CATH database. They described instances spanning most conceivable scenarios, ranging from those where fine mechanistic details are preserved even when sequence similarity is low to examples where obviously homologous proteins have essentially no functional similarity. Devos & Valencia (26) investigated the similarity between enzyme class, functional descriptions from key words, cellular functional classes, and binding sites in proteins with similar structures showing significant (but low) sequence similarity. They found that these attributes can only be reliably transferred from one protein to another to a limited degree and furthermore established sequence identity limits for such transfers. These studies suggest that one must exercise caution when attempting to transfer functional information between proteins, particularly when sequence identities are low (i.e., $<40\%$).

Approaches have also been developed to identify functionally important regions on protein surfaces (e.g., 1, 18, 57, 59) and to predict when functional sites on one protein can be used to predict a similar site on others (1). Although it is important to recognize similarities in functional sites, differences are also important because they are often related to specificity. Many protein families share details of molecular function (e.g., dehydrogenase) but vary in finer details such as their particular substrates (e.g., lactate/malate) or interacting partners. Methods have thus been developed that attempt to identify not just common functional residues but those

that are important in discerning subclasses (e.g., 18, 40, 57, 59). For example, key catalytic residues are common to all protein kinases, but two regions of the sequence differ and are known to confer the specificity for either serine/threonine (which differ only slightly) or the chemically distinct tyrosine. Such methods will be vital for the future development of evolution-based hierarchies of domain subfamilies, families, and superfamilies.

Using Domains to Interpret the Pathoetiology of Disease

Understanding domain function can lead to a greater appreciation of organismal function, particularly for domains present in the products of genes mutated in human disease (disease genes). Sequence analysis in general, and domain detection in particular, has played a pivotal role in elucidating the pathoetiology of disease (105). The majority (91%) of human disease gene products contain a domain assigned by Pfam or SMART (35). Thus both resources should represent the first ports-of-call for investigators wishing to predict the functions of newly sequenced disease genes.

The identification of disease gene orthologs and paralogs, by sequence analysis and domain prediction, also plays a part in understanding the molecular bases of diseases. Ortholog prediction allows the investigation of gene function in model organisms such as mouse, fruit fly, and nematode worm. Identification of human paralogs of disease genes generates additional disease gene candidates; this is because paralogs are often mutated in similar diseases. Due care is required, however, to establish that orthologs contain identical domain compositions and orders and to conserve functionally important active and/or binding sites. Otherwise, as was described previously for Ky orthologs (Figure 5), the conservation of function becomes less certain.

Although the molecular and/or domain functions of most newly identified disease genes can be predicted by sequence analysis, this knowledge is often inadequate to establish the critical deficit in organismal function for individuals suffering from the disease. If this divide between genotype and phenotype is to be spanned, protein function must be deduced from its constituent domains' functions; cellular function must be established from determination of the set of molecular interactions and reactions involving the protein; and organismal function must be considered as the grand synthesis of tissue expression, gene expression, post-translational modification, cellular localization, and molecular interaction data. The central position of domains within this informational hierarchy argues that their future contribution to biology will be enduring.

CONCLUSIONS

Many analogies can be drawn between the evolution of organisms and of domains. Competition among species and among domain families drives the selection of advantageous mutations in the occupation of both novel ecological and physiological

niches. Convergence of function is apparent for both species and domain family lineages. For example, the skull and body morphologies of the extinct marsupial wolf (Tasmanian tiger, *Thylacinus cynocephalus*) are similar to those of placental wolves (111), while trypsin-like and subtilisin-like serine proteases have independently evolved equivalent active sites. The embedding of domains within other domains parallels the presence of cellular endosymbionts within other cells: The insertion of SH2 and SH3 domains within a PH domain that is inserted within a TIM barrel domain (70) mirrors the presence of endosymbionts within endosymbionts within insect cells (116). Horizontal gene transfer between organisms is comparable to the transfer of genetic material between domain families which gives rise to sequence- and structure-similar motifs in unrelated molecules, as described previously for HhH and Asp box motifs. Indeed, the idea that domains once were composed mostly of multiple polypeptide segments (ADSs) is the molecular counterpart to the proposition that prokaryotes' genomes have a composite origin.

These analogies extend to the systems of classifying organisms and domains. In 1735 Carl Linnaeus published his *Systema Naturae*, a species classification system, which was codified one hundred years later (108). Modern domain classification systems are analogous to the Linnean system, not only because they are based on predicted evolutionary relationships but also because they are equivalent in structure. Just as a taxonomic family may contain several genera (a genus is a group of closely related species), a domain family also may contain several distinct sets of related domains (subfamilies). Progressing toward the base of the taxonomic tree from family to order to class to phylum is analogous to domain-based classification schema [e.g., SCOP (60)] that proceed from family to superfamily to fold to class.

For over a century zoologists have classified organisms using the Linnean system in order to provide insights into their natural history. Biologists are beginning to appreciate the benefits of hierarchical domain classification systems based on sequence, structure, and evolution. The numerous parallels between these systems suggest that domain classifications will prove to be key to our further understanding of the natural history of domain families.

ACKNOWLEDGMENT

We would like to thank Prof. A. N. Lupas for helpful discussions.

LITERATURE CITED

1. Aloy P, Querol E, Aviles FX, Sternberg MJE. 2001. Automated structure-based prediction of functional sites in proteins: applications to assessing the validity of inheriting protein function from homology in genome annotation and to protein docking. *J. Mol. Biol.* 311:395–408
2. Altschul SF, Boguski MS, Gish W,

Wootton JC. 1994. Issues in searching molecular sequence databases. *Nat. Genet.* 6:119–29
3. Andrade MA, Perez-Iratxeta C, Ponting CP. 2001. Protein repeats: structures, functions and evolution. *J. Struct. Biol.* 134:117–31
4. Apweiler R, Attwood TK, Bairoch A, Bateman A, Birney E, et al. 2000. InterPro—an integrated documentation resource for protein families, domains and functional sites. *Bioinformatics* 16:1145–50
5. Aravind L, Landsman D. 1998. AT-hook motifs identified in a wide variety of DNA-binding proteins. *Nucleic Acids Res.* 26:4413–21
6. Aravind L, Subramanian G. 1999. Origin of multicellular eukaryotes—insights from proteome comparisons. *Curr. Opin. Genet. Dev.* 9:688–94
7. Artymiuk PJ, Poirette AR, Rice DW, Willet P. 1997. A polymerase palm domain in adenylyl cyclase? *Nature* 388:33–34
8. Bateman A, Birney E, Durbin R, Eddy SR, Howe KL, Sonnhammer EL. 2000. The Pfam protein families database. *Nucleic Acids Res.* 28:263–66
9. Beckmann G, Hanke J, Bork P, Reich JG. 1998. Merging extracellular domains: fold prediction for laminin G-like and amino-terminal thrombospondin-like modules based on homology to pentraxins. *J. Mol. Biol.* 275:725–30
10. Beebe KD, Shin J, Peng J, Chaudhury C, Khera J, Pei D. 2000. Substrate recognition through a PDZ domain in tail-specific protease. *Biochemistry* 39:3149–55
11. Betts MJ, Guigo R, Agarwal P, Russell RB. 2001. Exon/intron structure conservation in the absence of protein sequence similarity: a record of dramatic events in evolution? *EMBO J.* 20:5354–60
12. Blake CCF. 1978. Do genes-in-pieces imply proteins-in-pieces? *Nature* 273:267
13. Blake CCF, Koenig DF, Mair GA, North ACT, Phillips DC, Sarma VR. 1965. Structure of hen egg-white lysozyme. *Nature* 206:757–61
14. Blanco G, Coulton GR, Biggin A, Grainge C, Moss J, Barrett M, et al. 2001. The kyphoscoliosis (ky) mouse is deficient in hypertrophic responses and is caused by a mutation in a novel muscle-specific protein. *Hum. Mol. Genet.* 10:9–16
15. Brannigan JA, Dodson G, Duggleby HJ, Moody PC, Smith JL, et al. 1995. A protein catalytic framework with an N-terminal nucleophile is capable of self-activation. *Nature* 378:416–19
16. Bystroff C, Baker D. 1998. Prediction of local structure in proteins using a library of sequence-structure motifs. *J. Mol. Biol.* 281:565–77
17. Bystroff C, Thorsson V, Baker D. 2000. HMMSTR: a hidden Markov model for local sequence-structure correlations in proteins. *J. Mol. Biol.* 301:173–90
18. Casari G, Sander C, Valencia A. 1995. A method to predict functional residues in proteins. *Nat. Struct. Biol.* 2:171–78
19. Cavalier-Smith T. 1985. Selfish DNA and the origin of introns. *Nature* 315:283–84
20. Chervitz SA, Aravind L, Sherlock G, Ball CA, Koonin EV, et al. 1998. Comparison of the complete protein sets of worm and yeast: orthology and divergence. *Science* 282:2022–28
21. Copley RR, Bork P. 2000. Homology among $(\beta\alpha)_8$ barrels: implications for the evolution of metabolic pathways. *J. Mol. Biol.* 303:627–41
22. Copley RR, Russell RB, Ponting CP. 2001. Sialidase like Asp-boxes: sequence-similar structures within different protein folds. *Protein Sci.* 10:285–92
23. Copley RR, Schultz J, Ponting CP, Bork P. 1999. Protein families in multicellular organisms. *Curr. Opin. Struct. Biol.* 9:408–15
24. Deshimaru M, Ogawa T, Nakashima K, Nobuhisa I, Chijiwa T, et al. 1996. Accelerated evolution of *crotalinae* snake venom gland serine proteases. *FEBS Lett.* 397:83–88

25. de Souza SJ, Long M, Klein RJ, Roy S, Lin S, Gilbert W. 1998. Toward a resolution of the introns early/late debate: Only phase zero introns are correlated with the structure of ancient proteins. *Proc. Natl. Acad. Sci. USA* 95:5094–99
26. Devos D, Valencia A. 2000. Practical limits of function prediction. *Proteins* 41:98–107
27. Dodson G, Wlodawer A. 1998. Catalytic triads and their relatives. *Trends Biochem. Sci.* 23:347–52
28. Doherty AJ, Serpell LC, Ponting CP. 1996. The helix-hairpin-helix DNA-binding motif: a structural basis for non-sequence-specific recognition of DNA. *Nucleic Acids Res.* 24:2488–97
29. Doolittle WF. 1978. Genes in pieces: Were they ever together? *Nature* 272:581–82
30. Doolittle WF, Brown JR. 1994. Tempo, mode, the progenote, and the universal root. *Proc. Natl. Acad. Sci. USA* 91:6721–28
31. Garcia-Vallve S, Romeu A, Palau J. 2000. Horizontal gene transfer in bacterial and archaeal complete genomes. *Genome Res.* 10:1719–25
32. Gerstein M. 1997. A structural census of genomes: comparing bacterial, eukaryotic, and archaeal genomes in terms of protein structure. *J. Mol. Biol.* 274:562–76
33. Gilbert W. 1978. Why genes in pieces? *Nature* 271:501
34. Gilbert W. 1986. The RNA world. *Nature* 319:618
35. Goodstadt L, Ponting CP. 2001. Sequence variation and disease in the wake of the draft human genome. *Hum. Mol. Genet.* 10:2209–14
36. Gould SJ. 1994. Tempo and mode in the macroevolutionary reconstruction of Darwinism. *Proc. Natl. Acad. Sci. USA* 91:6764–71
37. Grishin NV. 2001. Fold change in the evolution of protein structure. *J. Struct. Biol.* 134:167–85
38. Hadley C, Jones DT. 1999. A systematic comparison of protein structure classifications: SCOP, CATH and FSSP. *Structure* 7:1099–112
39. Haft DH, Loftus BJ, Richardson DL, Yang F, Eisen JA, et al. 2001. TIGRFAMs: a protein family resource for the functional identification of proteins. *Nucleic Acids Res.* 29:41–43
40. Hannenhalli SS, Russell RB. 2000. Analysis and prediction of protein functional sub-types from multiple sequence alignments. *J. Mol. Biol.* 303:61–76
41. Hofmann K. 2000. Sensitive protein comparisons with profiles and hidden Markov models. *Brief. Bioinform.* 1:167–78
42. Hofmann K, Bucher P, Falquet L, Bairoch A. 1999. The PROSITE database, its status in 1999. *Nucleic Acids Res.* 27:215–19
43. Holm L, Sander C. 1994. Structural similarity of plant chitinase and lysozymes from animals and phage: an evolutionary connection. *FEBS Lett.* 340:129–32
44. Holm L, Sander C. 1996. Mapping the protein universe. *Science* 273:595–603
45. Holm L, Sander C. 1997. Decision support system for the evolutionary classification of protein structures. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* 5:140–46
46. Holm L, Sander C. 1997. Enzyme HIT. *Trends Biochem. Sci.* 22:116–17
47. Holm L, Sander C. 1998. Dictionary of recurrent domains in protein structures. *Proteins* 33:88–96
48. Islam SA, Luo J, Sternberg MJE. 1995. Identification and analysis of domains in proteins. *Protein Eng.* 8:513–25
49. Jacq B. 2001. Protein function from the perspective of molecular interactions and genetic networks. *Brief. Bioinform.* 2:38–50
50. Ji Y, Schroeder D, Byrne D, Zipser B, Jellies J, et al. 2001. Molecular identification and sequence analysis of Hillarin, a novel protein localized at the axon hillock. *Biochim. Biophys. Acta* 1519:246–49
51. Kartha G, Bello J, Harker D. 1967.

Tertiary structure of ribonuclease. *Nature* 213:862–65
52. Koonin EV. 2001. An apology for orthologs—or brave new memes. *Gen. Biol.* 2:COMMENT1005
53. Koretke KK, Lupas AN, Warren PV, Rosenberg M, Brown JR. 2000. Evolution of two-component signal transduction. *Mol. Biol. Evol.* 17:1956–70
54. Kraulis PJ. 1991. MOLSCRIPT: a program to produce both detailed and schematic plots of protein structures. *J. Appl. Crystallogr.* 24:946–50
55. Kurland CG. 2000. Something for everyone. Horizontal gene transfer in evolution. *EMBO Rep.* 1:92–95
56. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, et al. 2001. Initial sequencing and analysis of the human genome. *Nature* 409:860–921. Errata. 2001. *Nature* 412:565–66
57. Landgraf R, Xenarios I, Eisenberg D. 2001. Three-dimensional cluster analysis identifies interfaces and functional residue clusters in proteins. *J. Mol. Biol.* 307:1487–502
58. Leonard CJ, Aravind L, Koonin EV. 1998. Novel families of putative protein kinases in bacteria and archaea: evolution of the "eukaryotic" protein kinase superfamily. *Genome Res.* 8:1038–47
59. Lichtarge O, Bourne HR, Cohen FA. 1996. An evolutionary trace method defines binding surfaces common to protein families. *J. Mol. Biol.* 257:342–58
60. Lo Conte L, Ailey B, Hubbard TJ, Brenner SE, Murzin AG, Chothia C. 2000. SCOP: a structural classification of proteins database. *Nucleic Acids Res.* 28:257–59
61. Lupas AN, Ponting CP, Russell RB. 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
62. MacCallum RM, Kelley LA, Sternberg MJ. 2000. SAWTED: structure assignment with text description—enhanced detection of remote homologous with automated SWISS-PROT annotation comparison. *Bioinformatics* 16:125–29
63. Makarova KS, Grishin NV. 1999. Thermolysin and mitochondrial processing peptidase: how far structure-functional convergence goes. *Protein Sci.* 8:2537–40
64. Matsuo Y, Bryant SH. 1999. Identification of homologous core structures. *Proteins* 35:70–79
65. Murzin AG. 1992. Structural principles for the propeller assembly of β-sheets: the preference for seven-fold symmetry. *Proteins* 14:191–201
66. Murzin AG. 1993. Sweet-tasting protein monellin is related to the cystatin family of thiol proteinase inhibitors *J. Mol. Biol.* 230:689–94
67. Murzin AG. 1995. A ribosomal protein module in EF-G and DNA gyrase. *Nat. Struct. Biol.* 2:25–26
68. Murzin AG. 1998. Probable circular permutation in the flavin-binding domain. *Nat. Struct. Biol.* 5:101
69. Murzin AG, Brenner SE, Hubbard T, Chothia C. 1995. SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* 247:536–40
70. Musacchio A, Gibson T, Rice P, Thompson J, Saraste M. 1993. The PH domain: a common piece in the structural patchwork of signalling proteins. *Trends Biochem. Sci.* 18:343–48
71. Nakai K. 2000. Protein sorting signals and prediction of subcellular localization. *Adv. Protein Chem.* 54:277–344
72. Ochman H, Lawrence JG, Groisman EA. 2000. Bacterial gene transfer and the nature of bacterial innovation. *Nature* 405:299–304
73. Ochman H, Moran NA. 2001. Genes lost and genes found: evolution of bacterial pathogenesis and symbiosis. *Science* 292:1096–98
74. Ohno S. 1999. Gene duplication and the

uniqueness of vertebrate genomes circa 1970–1999. *Semin. Cell Dev. Biol.* 10: 517–22
75. Orengo CA, Jones DT, Thornton JM. 1994. Protein superfamilies and domain superfolds. *Nature* 372:631–34
76. Orengo CA, Michie AD, Jones S, Jones DT, Swindells MB, Thornton JM. 1997. CATH—a hierarchic classification of protein domain structures. *Structure* 5:1093–108
77. Orgel LE, Crick FH. 1980. Selfish DNA: the ultimate parasite. *Nature* 284:604–7
78. Park J, Karplus K, Barrett C, Hughey R, Haussler D, et al. 1998. Sequence comparisons using multiple sequences detect three times as many remote homologues as pairwise methods. *J. Mol. Biol.* 284:1201–10
79. Patthy L. 1994. Exons and introns. *Curr. Opin. Struct. Biol.* 4:383–92
80. Patthy L. 1998. Genome evolution and the evolution of exon-shuffling—a review. *Gene* 238:103–14
81. Patthy L. 1999. *Protein Evolution*. Oxford: Blackwell Sci.
82. Ponting CP. 1997. Evidence for PDZ domains in bacteria, yeast, and plants. *Protein Sci.* 6:464–68
83. Ponting CP. 1999. Chlamydial homologues of the MACPF (MAC/perforin) domain. *Curr. Biol.* 9:R911–13
84. Ponting CP, Aravind L, Schultz J, Bork P, Koonin EV. 1999. Eukaryotic signalling domain homologues in archaea and bacteria. Ancient ancestry and horizontal gene transfer. *J. Mol. Biol.* 289:729–45
85. Ponting CP, Pallen MJ. 1999. β-propeller repeats and a PDZ domain in the tricorn protease: predicted self-compartmentalisation and C-terminal polypeptide-binding strategies of substrate selection. *FEMS Microbiol. Lett.* 179: 447–51
86. Ponting CP, Pallen MJ. 1999. A β-propeller within TolB. *Mol. Microbiol.* 31: 739–40
87. Ponting CP, Russell RB. 2000. Identification of distant homologues of fibroblast growth factors suggests a common ancestor for all β-trefoil proteins. *J. Mol. Biol.* 302:1041–47
88. Ponting CP, Schultz J, Copley RR, Andrade MA, Bork P. 2000. Evolution of domain families. *Adv. Protein Chem.* 54:185–244
89. Ray M-C, Germon P, Vianney A, Portalier R, Lazzaroni JC. 2000. Identification by genetic suppression of *Escherichia coli* TolB residues important for TolB-Pal interaction. *J. Bacteriol.* 182:821–24
90. Rossmann MG, Moras D, Olsen KW. 1974. Chemical and biological evolution of nucleotide-binding protein. *Nature* 250:194–99
91. Russell RB. 1994. Domain insertion. *Protein Eng.* 7:1407–10
92. Russell RB. 1998. Detection of protein three-dimensional side-chain patterns: new examples of convergent evolution. *J. Mol. Biol.* 279:1211–27
93. Russell RB, Ponting CP. 1998. Protein fold irregularities that hinder sequence analysis. *Curr. Opin. Struct. Biol.* 8:364–71
94. Russell RB, Saqi MA, Sayle RA, Bates PA, Sternberg MJ. 1997. Recognition of analogous and homologous protein folds: analysis of sequence and structure conservation *J. Mol. Biol.* 269:423–39
95. Russell RB, Saseini PD, Sternberg MJ. 1998. Supersites within superfolds: binding site similarity in the absence of homology. *J. Mol. Biol.* 282:903–18
96. Schäffer AA, Aravind L, Madden TL, Shavirin S, Spouge JL, et al. 2001. Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements. *Nucleic Acids Res.* 29:2994–3005
97. Schultz J, Milpetz F, Bork P, Ponting CP. 1998. SMART, a simple modular architecture research tool: identification of signaling domains. *Proc. Natl. Acad. Sci. USA* 95:5857–64

98. Shapiro L, Scherer PE. 1998. The crystal structure of a complement-1q family protein suggests an evolutionary link to tumor necrosis factor. *Curr. Biol.* 8:335–38
99. Sharp PA. 1985. On the origin of RNA splicing and introns. *Cell* 42:397–400
100. Shaw G. 1996. The pleckstrin homology domain: an intriguing multifunctional protein module. *Bioessays* 18:35–46
101. Siddiqui AS, Barton GJ. 1994. Continuous and discontinuous domains: an algorithm for the automatic generation of reliable protein domain definitions. *Protein Sci.* 4:872–84
102. Silverstein KA, Kilian A, Freeman JL, Johnson JE, Awad IA, Retzel EF. 2000. PANAL: an integrated resource for protein sequence ANALysis. *Bioinformatics* 16:1157–58
103. Simons KT, Ruczinski I, Kooperberg C, Fox BA, Bystroff C, Baker D. 1999. Improved recognition of native-like protein structures using a combination of sequence-dependent and sequence-independent features of proteins. *Proteins* 34:82–95
104. Sowdhamini R, Blundell TL. 1995. An automatic method involving cluster analysis of secondary structures for the identification of domains in proteins. *Protein Sci.* 4:506–20
105. Sreekumar KR, Aravind L, Koonin EV. 2001. Computational analysis of human disease-associated genes and their protein products. *Curr. Opin. Genet. Dev.* 11:247–57
106. Stanhope MJ, Lupas A, Italia MJ, Koretke KK, Volker C, Brown JR. 2001. Phylogenetic analyses do not support horizontal gene transfers from bacteria to vertebrates. *Nature* 411:940–44
107. Stoltzfus A, Spencer DF, Zuker M, Logsdon JM, Doolittle WF. 1994. Testing the exon theory of genes: the evidence from protein structure. *Science* 265:202–7
108. Strickland HE, Philipps J, Richardson J, Owen R, Jenyns L, et al. 1843. Report of a committee appointed "to consider the rules by which the Nomenclature of Zoology may be established on a uniform and permanent basis." *Br. Assoc. Adv. Sci. Rep. 12th Meet.* pp. 105–21
109. Stuart DI, Levine M, Muirhead H, Stammers DK. 1979. Crystal structure of cat muscle pyruvate kinase at a resolution of 2.6 Å. *J. Mol. Biol.* 134:109–42
110. Swindells MB. 1995. A procedure for detecting structural domains in proteins. *Protein Sci.* 4:103–12
111. Thomas RH, Schaffner W, Wilson AC, Paabo S. 1989. DNA phylogeny of the extinct marsupial wolf. *Nature* 340:465–67
112. Todd AE, Orengo CA, Thornton JM. 2001. Evolution of function in protein superfamilies, from a structural perspective. *J. Mol. Biol.* 307:1113–43
113. Traut TW. 1988. Do exons code for structural or functional units in proteins? *Proc. Natl. Acad. Sci. USA* 85:2944–48
114. van Raaij MJ, Mitraki A, Lavigne G, Cusack S. 1999. A triple β-spiral in the adenovirus fibre shaft reveals a new structural motif for a fibrous protein. *Nature* 401:935–38
115. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, et al. 2001. The sequence of the human genome. *Science* 291:1304–51
116. von Dohlen CD, Kohler S, Alsop ST, McManus WR. 2001. Mealybug beta-proteobacterial endosymbionts contain gamma-proteobacterial symbionts. *Nature* 412:433–36
117. Wallace AC, Borkakoti N, Thornton JM. 1997. TESS: a geometric hashing algorithm for deriving 3D coordinate templates for searching structural databases. Application to enzyme active sites. *Protein Sci.* 6:2308–23
118. Weber CH, Vincenz C. 2001. The death domain superfamily: a tale of two interfaces? *Trends Biochem. Sci.* 26:475–81

119. Wolf YI, Brenner SE, Bash PA, Koonin EV. 1999. Distribution of protein folds in the three superkingdoms of life. *Genome Res.* 9:17–26
120. Wu G, Fiser A, ter Kuile B, Sali A, Muller M. 1999. Convergent evolution of *Trichomonas vaginalis* lactate dehydrogenase from malate dehydrogenase. *Proc. Natl. Acad. Sci. USA* 96:6285–90

Annual Review of Biophysics and Biomolecular Structure
Volume 31, 2002

CONTENTS