Journal of Structural Biology **134,** 191–203 (2001)
doi:10.1006/jsbi.2001.4393, available online at http://www.idealibrary.com on IDEAL®

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?

Andrei N. Lupas,*,[1] Chris P. Ponting,† and Robert B. Russell‡,[2]

*Bioinformatics, GlaxoSmithKline, UP1345, 1250 South Collegeville Road, Collegeville, Pennsylvania 19426-0989; †MRC Functional Genetics Unit, Department of Human Anatomy and Genetics, University of Oxford, South Parks Road, Oxford OX1 3QX, United Kingdom; and ‡EMBL, Meyerhofstrasse 1, D-69012 Heidelberg, Germany

Received March 30, 2001, and in revised form June 12, 2001

This paper presents and discusses evidence suggesting how the diversity of domain folds in existence today might have evolved from peptide ancestors. We apply a structure similarity detection method to detect instances where localized regions of different protein folds contain highly similar sequences and structures. Results of performing an all-on-all comparison of known structures are described and compared with other recently published findings. The numerous instances of local sequence and structure similarities within different protein folds, together with evidence from proteins containing sequence and structure repeats, argues in favor of the evolution of modern single polypeptide domains from ancient short peptide ancestors (antecedent domain segments (ADSs)). In this model, ancient protein structures were formed by self-assembling aggregates of short polypeptides. Subsequently, and perhaps concomitantly with the evolution of higher fidelity DNA replication and repair systems, single polypeptide domains arose from the fusion of ADSs genes. Thus modern protein domains may have a polyphyletic origin. © 2001 Academic Press

Key Words: **protein evolution; protein structure similarity; protein function.**

INTRODUCTION

> [Empedocles, ca. 440 BC] had a theory (somewhat fantastic, it must be admitted) of evolution and the survival of the fittest. Originally, "countless tribes of mortal creatures were scattered abroad endowed with all manner of forms, a wonder to behold." There were heads without necks, arms without shoulders, eyes without foreheads, solitary limbs seeking for union. These things joined together as each might chance; there were shambling creatures with countless hands, creatures with faces and breasts looking in different directions, creatures with the bodies of oxen and the faces of men, and others with the faces of oxen and the bodies of men. There were hermaphrodites combining the natures of men and women, but sterile. In the end, only certain forms survived. (Bertrand Russell, *A History of Western Philosophy*)

Proteins and their associated functions have evolved as a consequence of inherited alterations to genes. Thus, the huge spectrum of proteins observable today has its roots in genetic events operating on a set of ancestral genes, similar to the near endless complexity of language resulting from the operation of a set of grammatical rules on a limited vocabulary (e.g., Searls, 1997).

A major genetic event in the evolution of the protein world is *duplication,* whereby the whole or a part of a genome is duplicated by diverse means (Ohno, 1970). Intragenomic duplication is thought to have occurred on numerous occasions throughout evolution. These duplications have left their mark on the human genome as large-scale inter- and intrachromosomal similarities and as smaller scale DNA repeats (International Human Genome Sequencing Consortium, 2001). A related genetic event is *lateral transfer,* in which organisms (particularly prokaryotes) acquire parts of the genomes of other organisms (e.g., Ruepp *et al.,* 2000; Nelson *et al.,*

[1] Current address: Department of Protein Evolution, Max-Plank-Institute for Developmental Biology, Spemannstrasse 35, Tubingen, Germany.

[2] To whom correspondence and reprint requests should be addressed.

1999; Ponting *et al.,* 1999), frequently resulting in "imperfect duplications" of parts of the host genome. (Lateral transfer between organisms of the same species is called *conjugation* and represents the major sexual event in bacteria). A third genetic event in this class is the acquisition of genes by *infection* with viruses, plasmids, or other mobile genetic elements.

The gene complement of an organism, whether resulting from inheritance, duplication, lateral transfer, or infection, is constantly subjected to the most frequent of genetic events shaping protein evolution: *point mutations* and *insertions* and *deletions* (indels). These accumulate through random drift and natural selection (the "genetic clock"), leading to a gradual divergence of initially identical gene copies (called *homologs,* because they originated from the same ancestor). Where such divergence occurs between genes in different organisms (*orthologs*) it generally reflects the result of random drift, the two genes being under selective pressure to maintain the ancestral line. Where such divergence occurs between duplicated genes in the same organism, however, it generally reflects the adaptation of individual copies to separate (but frequently related) functions, resulting in the formation of *paralogs.* Such paralogy can occasionally lead to protein families with hundreds of members in the same organism, as seen, for example, in human G-protein-coupled receptors (Horn *et al.,* 2001). The diversity of paralogous families is further enhanced by *recombination,* resulting in chimeric forms with novel properties.

Gene fusion, often after a gene duplication event, also plays a major role in generating protein variants, as does the duplication of a gene portion encoding a single domain by *unequal recombination* (International Human Genome Sequencing Consortium, 2001). In extreme cases, these two types of genetic events result in the formation of gigantic proteins with hundreds of related domains, such as in the muscle proteins titin and nebulin. In terms of complexity, these two types of events are likely to have generated most functional innovation. Thus, although humans contain only about twice as many genes as fruit flies, the complexity of human proteins in terms of domain composition is substantially larger, helping to explain the difference in complexity between the two organisms.

These genetic events explain the likely origin of paralogous and multidomain proteins from an ancestral "vocabulary" of protein domains but do not address the more fundamental question of how the domains themselves, the building blocks of all proteins, arose. Protein domains are compact polypeptide structures, generally organized around a clearly recognizable hydrophobic core and associated with a specific function or activity. It is clear that they adopt only a limited number of folds (e.g., Chothia, 1991; Orengo *et al.,* 1994), yet it is unclear whether each fold originated just once (and propagated via divergent evolution) or on multiple occasions (convergent evolution of structures). It is equally unclear whether some seemingly different folds share a common ancestor or whether each arose separately in evolution. Certainly it has been recognized that some genetic events can lead to fundamental changes in the structure of protein domains whose genes are affected (e.g., Grishin, 2001b).

Foremost among these events is *circular permutation,* which presumably occurs by gene duplication, fusion, and partial deletion (e.g., Ponting and Russell, 1998) and which can lead to substantial changes in the topology of a protein fold. Evidence for past intragene duplications causing short repetitions within proteins and giving rise to new structural variants is also compelling (Andrade *et al.,* 2001; Kajava, 2001). Finally, *illegitimate recombination,* occurring between unrelated genes, also leads to new folds where the recombined parts prove structurally compatible. But one need not always look to such major events: indels, recognized as a major force in protein divergence, can lead to substantial and dramatic alterations of structure (e.g., Russell, 1994), such that some contemporary homologues do indeed appear to possess different folds (Grishin, 2001b).

This article discusses evidence that hints at the evolutionary origins of domains by considering the occurrence of structurally (and sometimes functionally) similar elements in seemingly different folds. These findings are only now coming to light as available sequence and structure data increase and as the sensitivity of protein sequence comparison methods improve. Consequently, it is only now that some very ancient genetic events are beginning to be detected. The findings discussed here suggest that domains in contemporary proteins may differ in two key respects from their ancient structural counterparts. First, where the structures of modern domains are single-chain, their ancient counterparts may have been oligomeric, formed from a conglomerate of short polypeptides. Second, it is possible that modern domains may be related not from a single evolutionary gene lineage, but rather from several. In other words, modern domains may not be monophyletic.

ON THE DETECTION OF MOTIFS THAT ARE SIMILAR IN SEQUENCE AND/OR STRUCTURE

The common ancestry (homology) of protein domains is usually inferred from similarities in sequence and/or structure, but often remains contentious since arguments are derived from present-day specimens, not from the fossil record. Advances in our ability to sequence DNA pieces from fossil samples (e.g., Pabo, 1993; Golenberg *et al.,* 1990) and to revive dormant spores, sometimes hundreds of millions of years old (Vreeland *et al.,* 2000), may help to change this in the future, but it should be recognized that many of the evolutionary events leading to the protein domains observable today occurred prior to the Last Common Ancestor and are thus well outside the reach of existing (or indeed imaginable) paleontological methods. In the foreseeable future the reconstruction of ancient events in protein evolution will therefore continue to be dependent on the analysis of similarities in sequence and structure.

Sequence-based methods for detecting potentially homologous proteins currently center on the use of the position-specific and iterative version of BLAST, PSI-BLAST (Altschul *et al.,* 1997). This algorithm provides robust confidence estimates for the biological relevance of protein sequence similarities. Use of PSI-BLAST, often simultaneously with other database search tools such as HMMER (hmmer.wustl.edu), has generated many predictions of evolutionary and structural similarities that later have been borne out by experiment.

It is also possible to infer a common ancestor by comparison of protein three-dimensional structures. It is well known that proteins can adopt similar structures in the absence of significant sequence identity. However, structure similarity alone is not necessarily sufficient to say confidently that two proteins share a common ancestor, since it is still unclear whether convergence has produced similar structures multiple times. Accordingly a number of methods have been developed which assess whether a structural similarity is likely to indicate a common evolutionary origin. Methods have considered unusual structural features (e.g., Murzin, 1995), structural similarity (e.g., Matsuo and Bryant, 1999), similarities in molecular function (particularly active sites), the degree of sequence similarity seen within a structure-based sequence alignment (Murzin, 1993; Russell *et al.,* 1997), or a combination of features (Holm and Sander, 1997).

Other methods focus not on similarity across entire domains, but on localized regions of sequence and/or structure. These studies have investigated both similar folds whose evolutionary relationships remain ambiguous and also different folds that might have evolved similar localized structures by convergence. Swindells (1993, 1994) exploited the fact that glycine frequently occupies normally forbidden parts of the Ramachandran φ/ϕ space to develop a method that can search for loop conformations using the main-chain dihedral angles. This lead to a functional classification of $\alpha\beta$ doubly wound nucleotide binding topologies and a previously unnoted functional similarity between the flavin mononucleotide and pyridoxal phosphate binding sites of flavodoxin and tryptophan synthetase (despite no other similarity in sequence or structure).

Russell (1998) performed an all-against-all comparison of known structures searching for common protein 3D[3] side-chain patterns within different protein folds. The search identified several similarities in structurally clustered side-chains that were not expected by chance alone. Importantly, this study did not use a requirement for these similarities to be arranged in sequence order. The majority of the examples that were found are likely to be the result of convergent evolution since the amino acids contained in the similar pattern occur in different orders along the polypeptide chain. The most well known example of this is the Ser/His/Asp catalytic triad of trypsin-like and subtilisin-like serine proteases (among others; e.g., Dodson and Wlodawer, 1998). What was unexpected was a handful of examples where different folds possessed regions of side-chain structural similarity that also showed main-chain similarity and occurred in a colinear fashion in a short stretch of the polypeptide chain. These cases were not at that time explicitly studied in detail. A subsequent modification of the method was used to search explicitly for serine protease inhibitor canonical loops (Jackson and Russell, 2000), which found several putative sites of protease inhibition or cleavage.

Inspired by the study above, Copley *et al.* (2001) applied a technique of Cα atom similarity searching like that of Swindells (1993) and Jackson and Russell (2000) to search for Asp-box motifs among known 3D structures. The method found ungapped stretches with a significantly small rmsd to a probe motif, where significance was assessed by a fit to an extreme value distribution. With this strategy they identified numerous additional Asp boxes from many different folds.

For the purposes of this review, we have further modified the above method (Russell 1998) and have undertaken an all-against-all comparison of all pro-

[3] Abbreviations used: 3D, three—dimensional; Ig, immunoglobulin; ADSs, antecedent domain segments; NMR, nuclear magnetic resonance; rmsd, root mean square deviation; PDB, Protein Data Bank; ATP, adenosine triphosphate; SCOP, structural classification of proteins; NCBI, National Center for Biotechnology Information; URL, universal resource locator. The standard one- and three-letter abbreviations for the amino acids and chemical element symbols for individual atoms are also used throughout.

teins of known structure. This search was designed to detect new instances of proteins with different folds that contain short stretches of amino acids with *both* similar sequences *and* 3D structures.

A Method for the Detection of Sequence- and Structure-Similar Motifs

Many of the details of the method have been described elsewhere (Russell, 1998). Given a pair of protein structures, the method first finds all possible sets of identical amino acids common to both structures, which are within interacting distance, and have similar interatomic distances. Residues are defined as being within interacting distance if key side-chain functional atoms (see Russell, 1998; for glycine, the Cα atoms are used) are within 12 Å. Groups of residues equivalenced between two structures were defined to have similar interatomic distances if the differences between Cα–Cα, Cβ–Cβ, and *functional–functional* (where functional atoms are defined as atoms on each side-chain that represent the approximate functional center; see Russell, 1998) distances were less then 7.5, 7.0, and 6.0 Å, respectively. Equivalenced groups of amino acids were then filtered by calculating a weighted rmsd and the associated statistical significance, *P* (see below).

For the purposes of this study, the method was modified in several ways. As previously, amino acids unlikely to be directly involved in molecular function were ignored. Here we defined this set of residues as those with only carbon and hydrogen in their side chains (Ala, Phe, Ile, Leu, Pro, and Val). Glycine was included since we wished to consider main-chain atoms that might be involved in function, and they are frequently involved in such interactions. We also made no requirement for residue conservation. The aim was to consider as many motifs as possible and not to exclude structures only because too few homologous sequences were available. Finally, we required that amino acids in any matched pattern occurred in the same order along the protein sequences and that they occurred within a stretch of 20 amino acids. We define these pairs of matched patterns as "motifs," akin to those found, for example, in PROSITE (Hofmann *et al.,* 1999). However, it is important to emphasize that these motifs are based not only on similarities in sequence, as in PROSITE, but also on similarities in structure (main-chain and side-chain conformations). From a sequence-only perspective, occurrence of some of the motifs may not be expected to be meaningful. However, the additional constraint of structural similarity adds statistical significance. For all motifs found, we calculated a weighted rmsd as described previously (Russell, 1998). Here, the method was also modified to include main-chain in addition to side-chain atoms. The statistical significance (*P* value) of any potential match was assessed as before by comparison to a set of randomly generated motifs. For this study, these random motifs were required to have the same sequence order and lie within a total sequence separation of 20 residues.

We performed an all-against-all comparison of representatives from the SCOP database (Murzin *et al.,* 1995). One representative was chosen from each *protein* division within SCOP. There were 39553 initial potential motifs with three or more residues having a probability $P \leq 10^{-6}$ (roughly the inverse of the number of pairwise comparisons performed) grouped into 750 by ignoring matches that were entirely contained within other, or duplicate matches from the same protein superfamilies. These matches were viewed interactively and were considered potentially interesting either if nonprotein atoms were bound near to the region in one or both of the structures or if the stretch was longer than six residues. Many motifs involved only single α-helices or β-turns, and there are possibly other potentially interesting matches, simply where no information on bound nonprotein atoms was available.

Motifs that were deemed of greatest interest after inspection are shown in Table I, with associated details. Note that some motifs were refined further by combining multiple hits from different members of the same protein superfamily; if multiple hits were found, only the common residue matches are shown. Multiple alignments of matched polypeptide regions were performed using STAMP (Russell and Barton, 1992), and all matches found were explored to see if the short stretch arose from a portion of a global structural similarity that was not described in SCOP.

Results of Applying the Method

Seven examples of sequence- and structure-similar motifs were found in different domain folds (Table I, Fig. 1). Most of these examples have been previously discussed in the literature, but at least one (the Fe-S binding site in trimethylamine dehydrogenase and 4Fe-4S ferredoxins) was, to our knowledge, not previously described. The list is by no means exhaustive. Owing to the limitations of the search method, motifs involving, for example, hydrophobic residue conservation (e.g., the HhH motifs; Doherty *et al.,* 1996) were not detected.

Cytochromes c. A common motif occurs in the large group of cytochromes c (Fig. 1A). Five different folds within the SCOP database were found to contain at least one CxxCH structure motif. This motif binds covalently to a heme group, via the two cysteines, and coordinates the bound iron with the his-

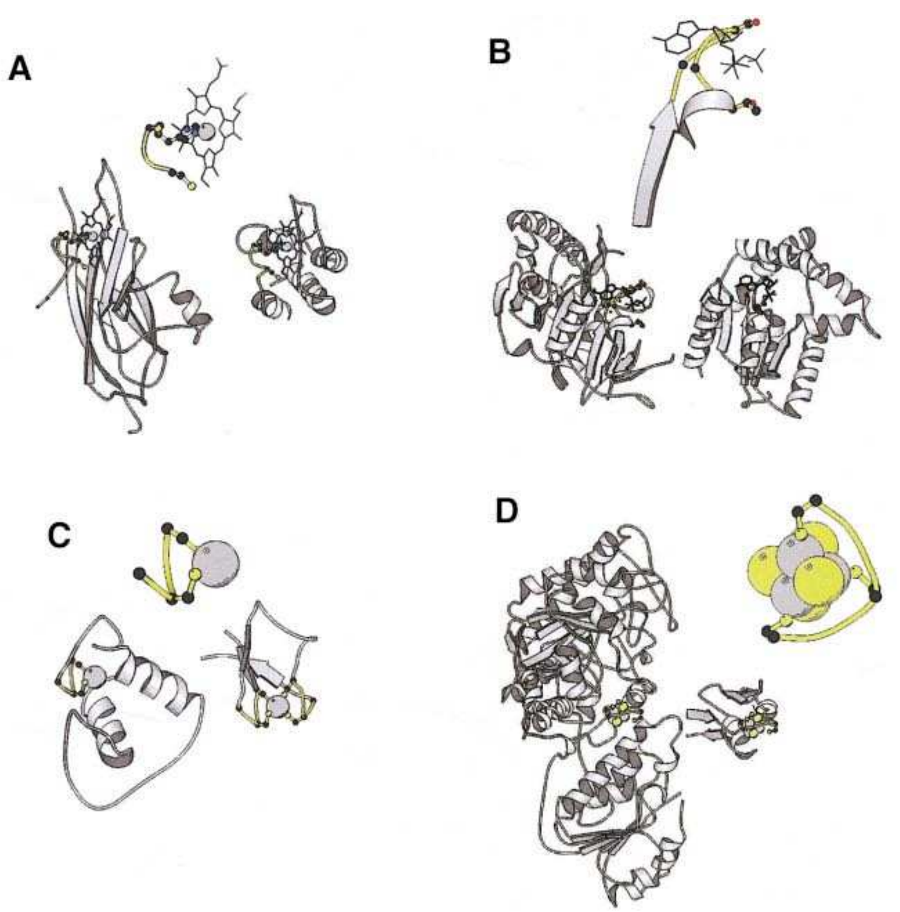
TABLE I
Motifs Found During This Study

	Pattern	Function	P value	SCOP	PDB	Range(s)
(i)	CxxCH	Heme attachment	10^{-7}–10^{-40}	S 26	2cy3	44–48 61–67 92–96 111–115
				α 3	1cor	12–16
				α 23	2ccy	118–122
				α 77	1prc	87–91 132–136 244–248 305–309
				β 2	1hcz	21–25
(ii)	CxxxxxRS	Ptase (types I/II)	10^{-7}–10^{-14}	$\alpha\beta$ 32	1vhr	124–131
				$\alpha\beta$ 31	1phr	12–19
(iii)	GxxGxxKT	P-loop	10^{-7}–10^{-25}	$\alpha\beta$ 71	1ayl	248–256
				$\alpha\beta$ 25	1ukz	23–31
(iv)	GxGxxG	FAD/NAD binding	10^{-8}–10^{-99}	$\alpha\beta$ 4	1gal	26–31
				$\alpha\beta$ 19	1xel	8–13
(v)	CxxCG	Zn finger	10^{-6}–10^{-99}	S 30	2nll	A135–139
				S 31	1aaf	15–19 36–40
				S 32	1zin	130–134 150–154
				S 33	1occ	F81–88
				S 38	1ptq	244–248 261–265
(vi)	CxxCxxC	Fe-S binding	10^{-7}	$\alpha\beta$ 18	2tmd	A345–351
				$\alpha + \beta$ 33	1vjw	10–16
(vii)	SxDGxxW	Asp box	10^{-7}–10^{-99}	β 1	1qba	845–852
				β 45	1eur	104–111 177–184 241–248 350–357

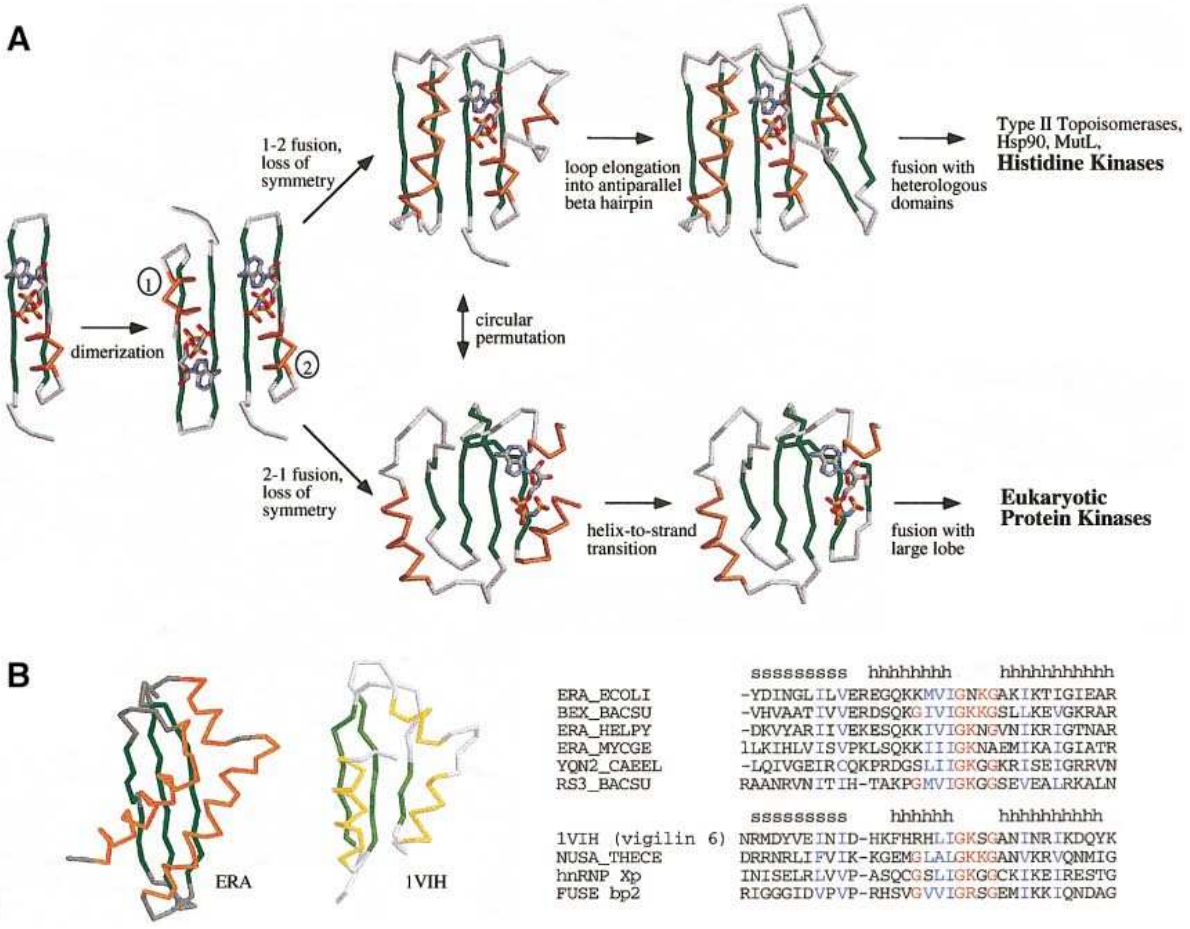
Note. Ranges of P values quoted are for all examples of the motif detected (i.e., there may be more than are shown). The Protein Data Bank codes are as follows. 2cy3, *Dsulfovibrio vulgaris* cytochrome c3; 1cor, *Pseudomonas stutzeri* cytochrome c551; 2ccy, *Rhodospirillum molishianum* cytochrome c′; 1prc, *Rhodopseudomonas viridus* photosynthetic reaction center; 1hcz, turnip cytochrome f; 1vhr, human phosphatase VHR; 1phr, bovine tyrosine phosphatase; 1ayl, *Escherichia coli* phosphoenolpyruvate carboxykinase; 1ukz, yeast uridylate kinase; 1gal, *Aspergillus niger* glucose oxidase; 1xel, *Escherichia coli* UDP-galactose 4-epimerase; 2nll, human retinoic acid receptor (DNA binding domain); 1aaf, HIV nucleocapsid protein; 1zin, *Bacillus stearothermophilus* adenylate kinase zinc finger domain; 1occ, bovine cytochrome c oxidase; 1ptq, mouse protein kinase C delta Cys2 domain; 2tmd, methylotropic bacterium trimethylamine dehydrogenase; 1vjw, *Thermotoga maritima* ferredoxin; 1qba, *Serratia marcescens* chitobiase; 1eur, *Micromonospora viridifaciens* neuraminidase.

tidine (Mathews, 1985). Within cytochrome c′ (2ccy), a single motif was found in the C-terminal helix of the four-helical bundle structure. Cytochrome c551 (1cor) and the many homologous cytochromes in this family contain one copy of the motif on the N-terminal helix of this all-α structure. In cytochrome f (1hcz), the motif lies in a C-terminal extension to an immunoglobulin (Ig) type β-sandwich structure. Within all of the cytochromes, the heme group functions in various electron transfer reactions. Sequence searches with this motif identified no sequences that were not known to belong to existing cytochrome c families.

Protein phosphatases. The catalytically essential cysteine and arginine residues of both type I and II phosphatases lie within the CxxxxxRS motif, which is contained within a loop connecting a β-strand to an α-helix around the active site of the enzymes. Despite apparently adopting different overall folds, the type I and type II protein phosphatases both contain this motif. Our inspection confirms a report that type I and type II phosphatases, partitioned into different folds in SCOP, are in fact related through a circular permutation (Fauman *et al.*, 1998; Grishin, 2001b). If the C-terminal residues 115–145 in the type II dual specificity phosphatases VHR (1vhr) are placed sequentially at the N-terminal end of the sequence (followed by residues 36–114) a good global alignment with the type I low-molecular-weight tyrosine phosphatase (1phr) can be obtained via structure comparison. The similarity involves four β-strands and three to four α-helices



1



```
                    sssssssss  hhhhhhh     hhhhhhhhhhh
ERA_ECOLI           -YDINGLILVEREGQKKMVIGNKGAKIKTIGIEAR
BEX_BACSU           -VHVAATIVVERDSQKGIVIGKKGSLLKEVGKRAR
ERA_HELPY           -DKVYARIIVEKESQKKIVIGKNGVNIKRIGTNAR
ERA_MYCGE           lLKIHLVISVPKLSQKKIIIGKNAEMIKAIGIATR
YQN2_CAEEL          -LQIVGEIRCQKPRDGSLIIGKGGKRISEIGRRVN
RS3_BACSU           RAANRVNITIH-TAKPGMVIGKGGSEVEALRKALN

                    sssssssss    hhhhhh    hhhhhhhhhh
1VIH (vigilin 6)    NRMDYVEINID-HKFHRHLIGKSGANINRIKDQYK
NUSA_THECE          DRRNRLIFVIK-KGEMGLALGKKGANVKRVQNMIG
hnRNP Xp            INISELRLVVP-ASQCGSLIGKGGCKIKEIRESTG
FUSE bp2            RIGGGIDVPVP-RHSVGVVIGRSGEMIKKIQNDAG
```

2

(depending on how lenient one is in assigning equivalences following structure comparison), and several additional conserved or semiconserved positions are apparent. The similarity means that it is likely that the type I and type II phosphatases share a common ancestor. One possibility is that tandem duplication of an ancestral phosphatase domain and subsequent N- and C- terminal truncation lead to a permuted variant by a mechanism that has been well described elsewhere (Ponting and Russell, 1995; Russell and Ponting 1998).

P-loops. Another frequently occurring motif that was found corresponds to the P-loops, which occur in many doubly wound $\alpha\beta$ structures, including Ras-p21 type GTPases (e.g., Walker *et al.,* 1982; Swindells, 1993, 1994) and phosphoenolpyruvate carboxy kinase (Fig. 1B), which adopts a different overall $\alpha\beta$ fold. This stretch of amino acids contains the motif GxxxGKT and functions by binding the phosphate backbone of a mononucleotide.

NAD/FAD binding motifs. There are also numerous similarities between NAD/FAD binding motifs within doubly wound $\alpha\beta$ proteins (Swindells, 1993). Many different folds within SCOP contain common GxGxxG motifs that adopt a conformation suited to bind the phosphate backbone of a dinucleotide. The majority of matches occur within doubly wound $\alpha\beta$ proteins and are generally involved in the binding of dinucleotides. Rossmann-like NAD(P) binding folds, trimethylamine dehydrogenase, FAD/NAD(P) binding domains, flavodoxin-like folds, isocitrate and isopropylate dehydrogenases, and phosphofructokinase are all examples of enzymes containing this motif, and in all the loop is involved in the interaction with a nucleotide phosphate backbone.

CxxCG zinc binding motifs. The CxxCG motif was found in a number of zinc-containing domains (Fig. 1C). Each of these domains is classified as a small, cysteine-rich domain, and their motifs are central to their core structures. Within rubredoxin-like proteins (e.g., the Bst ADK Zn domain from *B. stearothermophilus;* 1zin) two copies of this motif coordinate a single zinc and form the rubredoxin-like fold. Within the HIV-1 nucleocapsid protein (1aaf) there are two spatially separated copies of this motif. Each of these motifs packs against an HxxxxC motif, thus forming two similar distinct zinc binding domains. A similar picture is seen in the Cys2 ("C1") domain of protein kinase C d (1ptq), though here two CxxCG motifs coordinate a pair of zincs in conjunction with other parts of the structure that are quite different. Within the F subunit of cytochrome C oxidase (1occ) and the DNA binding domain of retinoic acid receptor, a single CxxCG motif coordinates a single zinc in conjunction with various other parts of the polypeptide chain.

4Fe-4S binding sites in ferredoxins and trimethylamine dehydrogenase. A striking Fe-S binding site common to a region of trimethylamine dehydrogenase (2tmd) and 4Fe-4S ferredoxins (e.g., ferredoxin A; 1vjw) is shown in Fig. 1D. The three cysteine (CxxCxxC) motif occurs within these two otherwise different protein structures and involves two turns of an α-helix, in addition to a four-residue loop, that surrounds a single iron–sulphur cluster, which is also coordinated by another cysteine from a different part of each structure (Cys 364 in 2tmd; Cys 51 in 1vjw). In the ferredoxin structure, the Fe-S cluster lies in a cleft formed by a pair of α-helices and a β-sheet; one of the helices is that involved in the match. Within the structure of trimethylamine dehydrogenase, the matched region occurs within a polypeptide segment linking the $\beta\alpha$ (TIM)-barrel domain and the middle ADP binding domain. In both proteins, the Fe-S cluster plays a role in electron transfer. The similarity between the Fe-S binding functions of these proteins has long been known (Hill *et al.,* 1977; Steenkamp and Singer, 1978), but to our knowledge no comments as to the main-chain structural similarity have been made.

Asp boxes. These sequence-similar and structure-similar motifs were found in the Ig-like domain of chitobiase and in bacterial sialidases in a previous

FIG. 1. Molscript (Kraulis, 1991) figures showing examples of similarities described in the text. (A) Cytochrome c heme attachment site from turnip cytochrome f (1hcz; left) and *Pseudomonas stutzeri* cytochrome c551 (1cor; right); (B) P-loop from *Escherichia coli* phosphoenolpyruvate carboxykinase (1ayl, left) and yeast uridylate kinase (1ukz, right); (C) Zn finger CxxCG motif from human retinoic acid receptor DNA binding domain (2nll, left) and *Bacillus stearothermophilus* adenylate kinase zinc finger domain (1zin, right); (D) Fe-S binding motif from methylotropic bacterium trimethylamine dehydrogenase (2tmd) and *Thermotoga maritima* ferredoxin (1vjw). For all examples, β-strands are shown as arrows, α-helices as ribbons, and all others regions as coils. Similar peptides are shown in yellow, with equivalent side-chains in ball-and-stick format and nonprotein atoms in stick form or as spheres (metal atoms).

FIG. 2. Further examples of structural similarity in seemingly unrelated proteins. (A) Hypothetical evolution of the ATP binding domains in bacterial histidine kinases and eukaryotic protein kinases from a nucleotide binding $\alpha\beta\beta$ element. (B) A comparison of the structures of the C-terminal domain of the bacterial GTPase ERA (PDB: 1EGA) and of a KH domain (shown here is the sixth domain of vigilin, PDB: 1VIH). Similarity is concentrated in two helices connected by an Ω loop, whose sequence is shown on the right. The underlying β-sheet has a different connectivity in the two structures.

study (Russell, 1998). This finding was corroborated by this modified approach and by a recent in-depth investigation of these "Asp-box" motifs in at least eight distinct protein families (Copley *et al.,* 2001). These motifs are found in at least three different folds: those of sialidases (a β-propeller) and chitobiase, as discussed above, and that of microbial ribonucleases (anti-parallel β-sheet with a single α-helix).

Other Examples

Two other examples not found in this search will be briefly described as they are informative to the discussion. Both are based on structures that are more recent than the database used in the search described above and have been partially discussed in the literature.

Bacterial histidine kinases and eukaryotic protein kinases. A recent search for proteins distantly related to the histidine kinase fold revealed a surprising topological similarity to the small lobe of eukaryotic protein kinases (Koretke *et al.,* 2000). The two folds bind nucleotides at equivalent sites and can be interrelated by circular permutation (Fig. 2A). Although they do not possess significant sequence similarity by the criteria applied to the other examples in this paper, their similarity was identified by a sequence search method, not by structural comparisons (Koretke *et al.,* 2000). Further inspection of the histidine kinase fold shows that it may be viewed as composed primarily of a duplicated $\alpha\beta\beta$ element, which encompasses much of the nucleotide binding site. Indeed, the two $\alpha\beta\beta$ elements can be superimposed with less than 2 Å rmsd in the backbone atoms. Thus, a hypothetical pathway may be constructed, which could account for the evolution of the ATP binding domains in bacterial histidine kinases and eukaryotic protein kinases from an ancestral nucleotide binding $\alpha\beta\beta$ peptide, using only established types of genetic events (Fig. 2A).

In this context it may be interesting to note that the structures of bacterial response regulators, which act as phosphatases of histidine kinases, can also be related by circular permutation to a large superfamily (HAD), which encompasses P type ATPases, phosphatases, epoxide hydrolases, and L-2-haloacid dehalogenases (Ridder and Dijkstra, 1999).

ERA and KH domains. A comparison of the C-terminal domain of the bacterial GTPase ERA with the KH domain, found in a large number of RNA binding proteins, revealed surprising similarities (Chen *et al.,* 1999), which appear correlated with the RNA binding function of the two domain types. Similarity is concentrated specifically in a long α-helix disrupted by an Ω turn, which also shows a similar pattern of glycine and lysine residues (Fig. 2B). Indeed, the similarity was noted by sequence comparisons prior to determination of the structure (Johnstone *et al.,* 1999) and prompted experiments that demonstrated RNA binding activity in ERA. Although the two domains also appear to be similar overall, the topology of the underlying β-sheet is different and cannot be interrelated by a single genetic event, such as circular permutation (Grishin, 2001a).

DISCUSSION

The Evolution of Sequence- and Structure-Similar Motifs in Different Folds

How did these sequence- and structure-similar motifs arise within apparently nonhomologous protein folds? There are four possible evolutionary scenarios for the evolution of these motifs that we consider here.

(i) First, the examples could all simply represent the coincident evolution of sequence and structure by *convergence* under functional constraints. That nature could have alighted more than once upon significantly similar sequences in similar structural arrangements cannot be discounted. One problem with this evolutionary route is that nature most often appears to evolve similar functions in unrelated protein families. Examples of active-site convergence that do *not* involve main-chain and sequence order conservation, such as the catalytic triads from serine proteases, are far more common than the sequence- and structure-similar motifs described here (Russell, 1998), and many proteins have similar functions with no molecular similarities at all (e.g., metal and serine proteases). It is very likely that different folds, once evolved, have converged to similar functions by more conventional evolutionary events, such as point mutations.

(ii) A second possibility is that sequence- and structure-similar motifs, present in different folds, have arisen by *divergence.* Vertical descent can lead to dramatic changes in structural topology, such that a global structural similarity is no longer apparent, despite similarities within highly conserved functional, such as active and binding, sites (Grishin, 2001b). Divergent evolutionary mechanisms, such as permutations, deletions, insertions, and rearrangements, might account for the different folds that contain a motif in common. Although this divergent evolutionary path might have been taken by proteins containing P-loops (example (iii)) or NAD/FAD binding motifs found in Rossmann-like doubly wound $\alpha\beta$ folds (example (iv)) there are no obvious divergent evolutionary mechanisms that might account for many of the observed fold differences.

(iii) A third evolutionary mechanism also involves divergent evolution and is analogous to the manner by which whole domains are thought to be shuffled into different protein contexts. This process is thought to be due to partial duplication of one gene and subsequent recombination within a domain-free region of a second gene. Rather than being duplicated and recombined into a gene region that does not encode a domain, these motifs might have been incorporated into domain-encoding gene regions, resulting in an insertion within the domain's tertiary structure. Thus, instead of being ancestral motifs about which folds are constructed (as above), these motifs are late additions that are grafted onto preexisting folds.

(iv) Although of the three evolutionary scenarios considered, this last is the most parsimonious (it involves the least number of evolutionary steps) it fails in one respect. This is that far from occurring on the periphery of their domain folds, these motifs can be deeply embedded and integral to their folds (Figs. 1 and 2) so they appear unlikely to have been grafted onto their folds by insertion. Thus we arrive at a fourth explanation for the evolution of sequence- and structure-similar motifs: that they represent the only remaining evidence of a predomain world, when protein structures consisted of conglomerations of short polypeptide chains [henceforth described as "antecedent domain segments" (ADSs)]. In this scenario the single-chain domains so prevalent now arose from the fusion of more ancient genes that encoded ADSs. The sequence- and structure-similar motifs are homologues that could have been found in many of these ancient domain conglomerates by dint of their significant contributions to molecular function. Their dispersal to domains of different folds arose from the supersedence of oligomeric domains by single-chain domains that came to be constructed from fusions of short genes (e.g., see Ponting and Russell, 2000). This explanation is similar to that given above (scenario ii) except that there is no requirement for those regions of domains that lie outside of their common motifs to be homologous.

Figure 3 shows an example of how two modern proteins, thioredoxin and disulfide bond forming protein (DsbA), might have formed from ADSs. Two minigene segments (red and blue in the figure) correspond to the ancient progenitors of a fused heterodimer that would be the common ancestor of both proteins. The possibility that such segments might have existed separately is supported by the observation that the two corresponding parts of thioredoxin reconsitute to form a folded and active protein (Slaby and Holmgren, 1975). Eventually the two separate ADSs are fused, and then are embellished in different ways to form thioredoxin (an N-terminal extension of a β-strand and an α-helix in white) and DsbA (a different N-terminal extension in yellow and the insertion of the green all-helical domain between the two ADSs).

Evidence for the Existence of ADSs from Repeats

The fourth scenario for the origins of these motifs is concordant with our expectation that significantly sequence-similar structures are homologous, having evolved divergently from a common ancestor. It also parallels, in many respects, a proposed description of the evolution of protein repeats (Ponting and Russell, 2000; Andrade *et al.,* 2001) described below.

Many single domain structures show a degree of internal symmetry. In many instances, these repetitive structures are associated with obvious sequence similarities. Folds such as horseshoe structures (e.g., ribonuclease inhibitor), single-stranded β-helices, those comprising ankyrin or HEAT repeats, and several others (see Andrade *et al.,* 2000, 2001; Marcotte *et al.,* 1999; Kajava, 2001) have clearly arisen by duplication of a single ancestral repeat, since the similarity between the repeats is detectable by standard sequence comparison techniques.

As discussed in Ponting and Russell (2000) and elsewhere in this volume (Andrade *et al.,* 2001), the evolution of repeat-containing domains is problematic since, if all repeats are homologous, they are all related by vertical descent to a single ancestral gene product that would be most unlikely to fold or perform a viable function, when in isolation. This is resolvable if the ancestral gene product formed homomultimers since these could associate to form structures equivalent to those of modern repeat assemblies.

The evolutionary scenario for the evolution of repeat assemblies suggests an age when domains were constructed from multiple *identical* chains. These multimeric domains would then be replaced gradually by single polypeptide chains encoding multiple repeats, since these are highly likely to be more efficiently folded and more thermodynamically stable. Folds such as those of immunoglobulins or ferredoxins, which contain a duplicated structure, may also have evolved by similar gene duplication events (for a nearly prescient view on this issue see McLachlan, 1972). Indeed, even domains not hitherto considered to contain internal duplications may be seen to contain cores that may have evolved from the duplication of a simpler supersecondary structure element (Fig. 2A). If so, the simpler multimeric precursors of these proteins are not observable to-

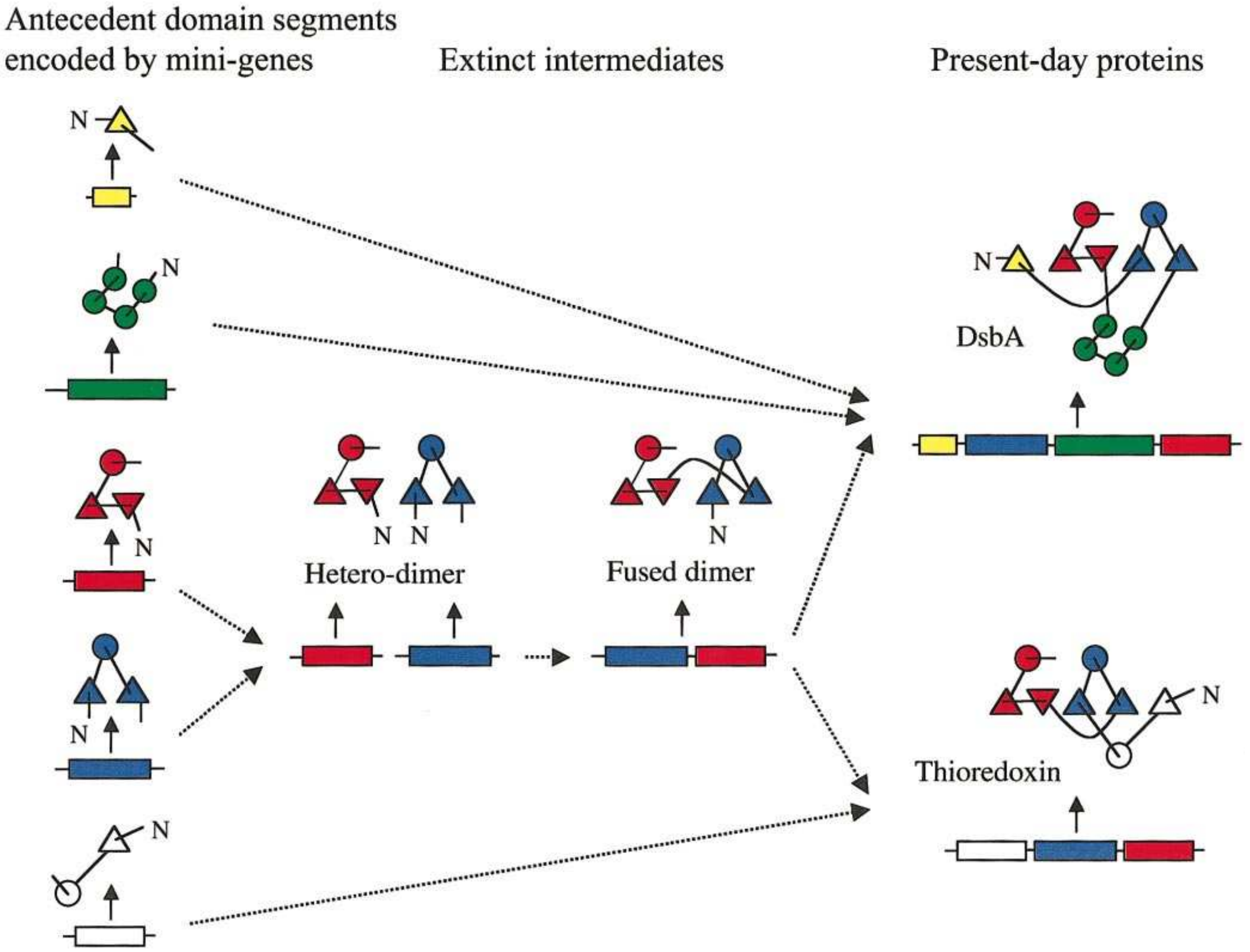


FIG. 3. Schematic showing the hypothetical evolutionary scenario that might have led to the evolution of thioredoxin and DsbA from a series of different antecedent domain segments (ADSs). α-Helices are represented by circles and β-strands by triangles. The different colored segments represent "minigenes" encoding the corresponding ADSs. Details are given in the text.

day. It may be argued from the structure of proteins found in Bacteria, Archaea, and eukaryotes that the age of these multimeric assemblies, if they existed, must have predated the Last Common Ancestor and therefore lie more than 3 billion years in our past.

If some ancient domains were composed of *homo*multimers, it is plausible that others may have been composed of *hetero*multimers. Again, since modern domains are not known to contain several chains, oligomeric domain ancestors must have been subsequently replaced by single-chain versions. That the conglomeration of multiple polypeptides can form a stable structure is amply shown by studies of the eukaryotic Cdc42/Rac interactive binding (CRIB) motif (Rudolph *et al.,* 1998), the G-protein γ-subunit (Sondek *et al.,* 1996), and a staphylococcal motif (Penkett *et al.,* 2000) that each lacks discernible tertiary structure in isolation, yet forms a stable structure upon binding another domain structure (a small GTPase, G-protein β-subunit, and fibronectin, respectively). Indeed, the proteolytic dissection of proteins such as Trp repressor or cytochrome c yields small fragments capable of undergoing spontaneous noncovalent association to form subdomains with native-like secondary and/or tertiary structural features (Wu *et al.,* 1994). These results suggest that protein domains contain autonomously folding supersecondary structure elements, which can reassemble into compact structures even when their polypeptide connections are severed.

A "mix-and-join" mechanism of the type suggested above probably occurred at several stages during evolution. Indeed, the observation that several multidomain proteins in certain organisms exist as separate chains in others (Enright *et al.,* 1999; Marcotte *et al.,* 1999) suggests that the mechanism has been used more recently in evolutionary history. It is difficult to ascertain estimates as to the number and nature of ancient ADSs, since evidence for them beyond a simple similarity in structure (i.e., function or structural constraints in the examples above) is rare. As more sequence and structure information becomes available, we anticipate that more examples will be uncovered providing a better picture as to whether a theory of such an ancient peptide world is tenable.

Some additional support for the theory comes, intriguingly, from recent improvements in protein tertiary structure prediction. Baker and co-workers (Bystroff *et al.,* 2000; Simons *et al.,* 1999; Bystroff and Baker, 1998) have achieved outstanding prediction success by methods that predict local regions of protein secondary and supersecondary structure. The methods make use of a library (I sites) that contains local regions of structure similarity common to different protein folds. It is tempting to suggest that some of these I sites may correspond to the ADSs discussed above. Thus prediction success could come about by identifying regions of genuine ancient homology between otherwise different three-dimensional structures.

The possibility of an early peptide world agrees with our preconceptions about the nature of early protein-based organisms. The sophisticated mechanisms of DNA replication that exist in nature today likely evolved from simpler systems that would have had higher error rates during DNA replication. The result would have been more mutations to protein sequences, making long protein sequences unfavorable, since they are more likely to contain mutations affecting folding. Errors in DNA replication also would have facilitated intragenome duplication resulting in multiple short polypeptide chain homologues.

The hypothesis also lends some support to the idea that short peptides played a role in the origin of life. According to the chemo-autotrophic theory, pyrite formation supplied the energy source for the nonbiological generation of peptides (Keller *et al.,* 1994). Experiments using such mineral surfaces are able to generate only short peptides of 10 amino acids or fewer (Ferris *et al.,* 1996). If the ADS hypothesis is correct then such short peptides may have had rudimentary functions and combinations of them could ultimately have lead to the first autocatalytic reproduction cycle.

CONCLUSIONS

The examples discussed here suggest the possibility of the evolution of complex protein folds from short peptides (antecedent domain segments, ADSs) via a series of gene duplication and fusion events. Many current folds may have arisen from multiple duplications of common peptide ancestors and, over time, lost most or all signals traditionally used to infer a common ancestor. Many others may have evolved from the fusion of multiple ADSs or from the illegitimate recombination of domains containing structurally compatible ADSs, thus being polyphyletic in origin.

The possibility that such polypeptides could function in isolation is easily testable by protein synthesis. Modified variants of either peptide repeats or short putative functional peptides common to different folds that can fold and function in isolation support the idea of a peptide world. Thus techniques traditionally associated with studying the physics and chemistry of protein structure and folding may begin to provide insights into the nature of early protein evolution.

REFERENCES

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

Andrade, M. A., Ponting, C. P., Gibson, T. J., and Bork, P. (2000) Homology-based method for identification of protein repeats using statistical significance estimates, *J. Mol. Biol.* **298,** 521–537.

Andrade, M. A., Perez-Iratxeta, C., and Ponting, C. P. (2001) Protein repeats: Structures, functions and evolution, *J. Struct. Biol.* **134,** 117–131.

Bystroff, C., Thorsson, V., and Baker, D. (2000) HMMSTR: A hidden Markov model for local sequence-structure correlations in proteins, *J. Mol. Biol.* **301,** 173–190.

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

Chen, X., Court, D. L., and Ji, X. (1999) Crystal structure of ERA: a GTPase-dependent cell cycle regulator containing an RNA binding motif, *Proc. Natl. Acad. Sci. USA* **96,** 8396–8401.

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

Dodson, G., and Wlodawer, A. (1998) Catalytic triads and their relatives, *Trends Biochem Sci.* **23,** 347–352.

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

Enright, A. J., Iliopoulos, Il, Kyrpides, N. C., and Ouzounis, C. A. (1999) Protein interaction maps for complete genomes based on gene fusion events, *Nature* **402,** 25–26.

Fauman, E. B., Cogswell, J. P., Lovejoy, B., Rocque, W. J., Holmes, W., Montana, V. G., Piwnica-Worms, H., Rink, M. J., and Saper, M. A. (1998) Crystal structure of the catalytic domain of the human cell cycle control phosphatase, Cdc25A, *Cell* **93,** 617–625.

Ferris, J. P., Hill, A. R. Jr., Liu, R., and Orgel, L. (1996) Synthesis of long prebiotic oligomers on mineral surfaces, *Nature* **381,** 20–21.

Golenberg, E. M., Giannasi, D. E., Clegg, M. T., Smiley, C. J., Durbin, M., Henderson, D., and Zurawski, G. (1990) Chloroplast DNA sequence from a miocene Magnolia species, *Nature* **344,** 656–658.

Grishin, N. V. (2001a) KH domain: One motif, two folds, *Nucleic Acids Res.* **29,** 638–643.

Grishin, N. V. (2001b) Fold change in the evolution of protein structure, *J. Struct. Biol.* **134,** 167–185.

Hill, C. L., Steenkamp, D. J., Holm, R. H., and Singer, T. P. (1977) Identification of the iron-sulfur center in trimethylamine dehydrogenase, *Proc. Natl. Acad. Sci. USA* **74,** 547–551.

Hofmann, K., Bucher, P., Falquet, L., and Bairoch, A. (1999) The PROSITE database, its status in 1999, *Nucleic Acids Res.* **27,** 215–219.

Holm, L., and Sander, C. (1997) Decision support system for the evolutionary classification of protein structures, *Proc. Int. Conf. Intell. Syst. Mol. Biol.* **5,** 140–146.

Horn, F., Vriend, G., and Cohen, F. E. (2001) Collecting and harvesting biological data: The GPCRDB and NucleaRDB information systems, *Nucleic Acids Res.* **29,** 346–349.

International Human Genome Sequencing Consortium (2001) Initial sequencing and analysis of the human genome. *Nature* **409,** 860–921.

Jackson, R. M., and Russell, R. B. (2000) The serine protease inhibitor canonical loop conformation: Examples found in extracellular hydrolases, toxins, cytokines and viral proteins, *J. Mol. Biol.* **296,** 325–334.

Johnstone, B. H., Handler, A. A., Chao, D. K., Nguyen, V., Smith, M., Ryu, S. Y., Simons, E. L., Anderson, P. E., and Simons, R. W. (1999) The widely conserved Era G-protein contains an RNA-binding domain required for Era function in vivo, *Mol. Microbiol.* **33,** 1118–1131.

Kajava, A. (2001) Proteins with repeated sequences: Structural prediction and modeling, *J. Struct. Biol.* **134,** 132–144.

Keller, M., Blochl, E., Wachetershauser, G., and Stetter, K. O. (1994) Formation of amide bonds without a condensation agent and implications for the origin of life, **368,** 836–838.

Koretke, K. K., Lupas, A. N., Warren, P. V., Rosenberg, M., and Brown, J. R. (2000) Evolution of two-component signal transduction, *Mol. Biol. Evol.* **17,** 1956–1970.

Kraulis, P. J. (1991) MOLSCRIPT: A program to produce both detailed and schematic plots of protein structures, *J. Appl. Crystallogr.* **24,** 946–950.

Marcotte, E. M., Pellegrini, M., Yeates, T. O., and Eisenberg, D (1999a) A census of protein repeats, *J. Mol. Biol.* **293,** 151–160.

Marcotte, E. M., Pellegrini, M., Thompson, J. J., Yeates, T. O., and Eisenberg, D. (1999b) A combined algorithm for genomewide prediction of protein function, *Nature* **402,** 83–86.

Mathews, F. S. (1985) The structure, function and evolution of cytochromes, *Prog. Biophys. Mol. Biol.* **45,** 1–56.

Matsuo, Y., and Bryant, S. H. (1999) Identification of homologous core structures, *Proteins* **35,** 70–79.

McLachlan, A. D. (1972) Repeating sequences and gene duplication in proteins, *J. Mol. Biol.* **64,** 417–437.

Murzin, A. G. (1992) Structural principles for the propeller assembly of beta-sheets: The preference for seven-fold symmetry, *Proteins* **14,** 191–201.

Murzin, A. G. (1993) Sweet-tasting protein monellin is related to the cystatin family of thiol proteinase inhibitors, *J. Mol. Biol.* **230,** 689–694.

Murzin, A. G. (1995) A ribosomal protein module in EF-G and DNA gyrase, *Nature Struct Biol.* **2,** 25–26.

Murzin, A. G., Lesk, A. M., and Chothia, C (1992) β-Trefoil fold: Patterns of structure and sequence in the Kunitz inhibitors interleukins-1 beta and 1 alpha and fibroblast growth factors, *J. Mol. Biol.* **223,** 521–543.

Murzin, A. G., Brenner, S. E., Hubbard, T., and Chothia, C. (1995) SCOP: A structural classification of proteins database for the investigation of sequences and structures, *J. Mol. Biol.* **247,** 536–540.

Nelson, K. E., Clayton, R. A., Gill, S. R., Gwinn, M. L., Dodson, R. J., Haft, D. H., Hickey, E. K., Peterson, J. D., Nelson, W. C., Ketchum, K. A., McDonald, L., Utterback, T. R., Malek, J. A., Linher, K. D., Garrett, M. M., Stewart, A. M., Cotton, M. D., Pratt, M. S., Phillips, C. A., Richardson, D., Heidelberg, J., Sutton, G. G., Fleischmann, R. D., Eisen, J. A., White, O., Salzberg, S. L., Smith, H. O., Venter, J. C., and Fraser, C. M. (1999) Evidence for lateral gene transfer between Archaea and Bacteria from genome sequence of *Thermotoga maritima, Nature* **399,** 323–329.

Ohno, S. (1970) Evolution by Gene Duplication, Springer-Verlag, Heidelberg, Germany.

Penkett, C. J., Dobson, C. M., Smith, L. J., Bright, J. R., Pickford, A. R., Campbell, I. D., and Potts, J. R. (2000) Identification of residues involved in the interaction of *Staphylococcus aureus* fibronectin-binding protein with the ^{4}F1^{5}F1 module pair of human fibronectin using heteronuclear NMR spectroscopy, *Biochemistry* **39,** 2887–2893.

Pabo, S. (1993) Ancient DNA, *Sci. Am.* **269,** 86–92.

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

Ponting, C. P., Aravind, L., Schultz, J., Bork, P., and Koonin, E. V. (1999) Eukaryotic signalling domain homologues in archaea and bacteria. Ancient ancestry and horizontal gene transfer, *J. Mol. Biol.* **289,** 729–745.

Ponting, C. P., and Russell, R. B. (1995) Swaposins: Circular permutations within genes encoding saposin homologues, *Trends Biochem. Sci.* **20,** 179–180.

Ridder, I. S., and Dijkstra, B. W. (1999) Identification of the Mg2+-binding site in the P-type ATPase and phosphatase members of the HAD (haloacid dehalogenase) superfamily by structural similarity to the response regulator protein CheY, *Biochem. J.* **339,** 223–226.

Rudolph, M. G., Bayer, P., Abo, A., Kuhlmann, J., Vetter, I. R., and Wittinghofer A. (1998) The Cdc42/Rac interactive binding region motif of the Wiskott Aldrich syndrome protein (WASP) is necessary but not sufficient for tight binding to Cdc42 and structure formation, *J. Biol. Chem.* **273,** 18067–18076.

Ruepp, A., Graml, W., Santos-Martinez, M. L., Koretke, K. K., Volker, C., Mewes, H. W. , Frishman, D., Stocker, S., Lupas, A. N., and Baumeister, W. (2000) The genome sequence of the thermoacidophilic scavenger *Thermoplasma acidophilum, Nature* **407,** 508–513.

Russell, R. B. (1998) Detection of protein three-dimensional sidechain patterns: New examples of convergent evolution, *J. Mol. Biol.* **279,** 1211–1227.

Russell, R. B. (1994) Domain insertion, *Protein Eng.* **7,** 1407–1410.

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

Russell, R. B., and Ponting, C. P. (1998) Protein fold irregularities that hinder sequence analysis, *Curr. Opin. Struct. Biol.* **8,** 364–371.

Saraste, M., Sibbald, P. R., and Wittinghofer, A. (1990) The P-loop—A common motif in ATP- and GTP-binding proteins, *Trends Biochem. Sci.* **15,** 430–434.

Searls, D. B. (1997) Linguistic approaches to biological sequences, *Comput. Appl. Biosci.* **13,** 333–344.

Simmons, K. T., Ruczinski, I., Kooperberg, C., Fox, B. A., Bystroff, C., and 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.

Slaby, I., and Holmgren, A. (1975) Reconstitution of Escherichia coli thioredoxin from complementing peptide fragments obtained by cleavage at methionine-37 or arginine-73, *J. Biol. Chem.* **250,** 1340–1347.

Sondek, J., Bohm, A., Lambright, D. G., Hamm, H. E., and Sigler, P. B. (1996) Crystal structure of a G-protein bg dimer at 2.1Å resolution, *Nature* **379,** 369–374.

Steenkamp, D. J., and Singer, T. P. (1978) Participation of the iron-sulphur cluster and of the covalently bound coenzyme of trimethylamine dehydrogenase in catalysis, *Biochem. J.* **169,** 361–369.

Swindells, M. B. (1993) Classification of doubly wound nucleotide binding topologies using automated loop searches, *Protein Sci.* **2,** 2146–2153.

Swindells, M. B. (1994) Loopy similarities, *Nature Struct. Biol.* **1,** 421–422.

Vreeland, R. H., Rosenzweig, W. D., and Powers, D. W. (2000) Isolation of a 250-million-year-old halotolerant bacterium from a primary salt crystal, *Nature* **407,** 897–900.

Walker, J. E., Saraste, M., Runswick, M. J., and Gay, N. J. (1982) Distantly related sequences in the alpha- and beta-subunits of ATP synthase, myosin, kinases and other ATP-requiring enzymes and a common nucleotide binding fold, *EMBO J.* **1,** 945–951.

Wallace, A. C., Borkakoti, N., and Thornton, J. M. (1997) TESS: A geometric hashing algorithm for deriving 3D coordinate templates for searching structural databases. Application to enzyme active sites, *Protein Sci.* **6,** 2308–2323.

Wallace, A. C., Laskowski, R. A., and Thornton, J. M. (1996) Derivation of 3D coordinate templates for searching structural databases: Application to Ser-His-Asp catalytic triads in the serine proteinases and lipases, *Protein Sci.* **5,** 1001–1013.

Wu, L. C., Grandori, R., and Carey, J. (1994) Autonomous subdomains in protein folding, *Protein Sci.* **3,** 369–371.