

THE MULTIPLICITY OF DOMAINS IN PROTEINS



Russell F. Doolittle

Center for Molecular Genetics University of California, San Diego, La Jolla,
California, 92093-0634

KEY WORDS: protein domains, domain evolution, modules, exon shuffling, evolution,
combinatorial advantage

CONTENTS

PERSPECTIVES AND SUMMARY	288
IDENTIFYING DOMAINS	289
Sequence Searching Programs	290
Three-Dimensional Comparison Programs	290
Classifying Domains	291
Nomenclature	292
EXONS, INTRONS, AND MODULES	292
Structural Aspects	293
Genetic Aspects of Domain Rearrangement	293
THE PHYLOGENETIC DISTRIBUTION OF DOMAINS	294
Some Ancient Domains	294
Intracellular Modular Proteins Apparently Unique to Eukaryotes	296
EXTRACELLULAR MOSAIC PROTEINS IN ANIMALS	297
The EGF-Like Domain	301
The Immunoglobulin Domain	301
The Fibronectin Type III Domain	302
Some Other Common Domains	303
MODULES UNIQUE TO PLANT OR FUNGAL PROTEINS	305
EVOLUTIONARILY MOBILE DOMAINS IN BACTERIA	306
SOME UNEXPECTED OCCURRENCES	308
HOW MANY DOMAINS?	308
SUMMARY AND CONCLUSIONS	309

ABSTRACT

The domainal nature of proteins is well established. What is less certain is
how many domains are evolutionarily mobile in that they occur in otherwise

287

0066-4154/95/0701-0287\$05.00

nonhomologous proteins or in different sequential locations in homologous proteins. The combinatorial advantage of shuffling domains around into diverse settings is obvious. Those domains that have been shuffled about in recent evolutionary times, within the last half billion years or so, can usually be identified on the basis of sequence resemblances alone. Contrarily, domains that were rearranged in ancient times may only be apparent after three-dimensional analysis, their sequence resemblances having been eroded over time. The shuffling of domains in recently evolved proteins has been greatly promoted by introns, but this does not imply that all domainal rearrangements involve introns. Only a small fraction of known exons show evidence of having been shuffled. Taken in aggregate, the available data best fit a scenario whereby a relative small number of genes encoding domain-sized polypeptides has been expanded by duplication and modification with a burst of exceptional genomic rearrangement.

PERSPECTIVES AND SUMMARY

It is generally accepted that there is a level of organization in proteins that overlaps the classical categories of tertiary and quaternary, i.e. sequentially consecutive residues in polypeptide chains tend to fold into more or less compact modules called domains (1). In this regard, a domain can be defined as that part of a protein that can fold up independently of neighboring sequences, although the term is often used less precisely. Crystallographers often interchange the term *domain* with the word *fold* (used as a noun). Sequence-oriented investigators often use the term domain when they mean sequence motif or signature sequence, although such sequences are often characteristic of domains.

The underlying basis of domain organization is the cohesion between sidechains that stabilizes unique structures. As such, there is a need for a critical number of residues in a sequence, typically 50–100, before a domain can be realized, although smaller segments can be further stabilized by folding around metal centers, or by forming appropriate disulfide bonds, or in the case of short repetitive units, packing against each other. Contrarily, some domains are much larger (2). As it happens, the average polypeptide chain length for proteins or their constituent subunits is about 350 residues (3), and most proteins contain two or more domains, although, naturally, many proteins are single domains.

The principal questions before us are, How many kinds of domains exist, and what was their origin? An extreme view is based on the idea that a large number of polypeptides was concocted early in the history of living systems and that all the world's proteins, past and present, are the result of shuffling those primordial structural units (4). An opposing view is that there were only a few small polypeptides in the early stages of life and that most contemporary

proteins are descended from them as a result of genetic duplication and descent with modification (5). Perhaps the truth will turn out to be somewhere between these two extremes, a modest number of starter types giving rise to the first genuinely modular proteins. We can attempt to evaluate these propositions retrospectively by considering the inventory of contemporary domains.

At the same time we can address some subsidiary questions, including: Are new domains still being evolved, and if so, how frequently? Also, does a similar fold always imply common ancestry? If we can conduct the census along phylogenetic grounds, perhaps we can obtain some idea of the general multiplicity. Thus, some domains are clearly ubiquitous and occur in all kinds of organisms. Accordingly, they must predate the last common ancestor of all contemporary life forms. Others appear to be narrowly distributed, occurring only in animals, for example. In some cases, this may be an illusion resulting from extensive sequence decay, and three-dimensional studies will be needed to settle the point.

The emphasis in this review is on domains that have been shuffled about during evolution, whether they were moved recently or in the very ancient past. The evolutionarily mobile domain is often called a module (as in *modular* furniture). Some authors prefer to limit the use of this term to domains that are genetically bordered by introns (6). In my view, this is too restrictive. By our criterion here, a domain need only occur in two different settings in otherwise nonhomologous proteins, or at clearly different locations in homologous proteins, to be considered evolutionarily mobile. Indeed, some domains have been so successful in their structural and functional adaptations that they occur in numerous proteins in very different settings.

IDENTIFYING DOMAINS

The domains in a given protein may be packed together tightly or loosely. If it is the latter, they can often be dissected by limited proteolysis, providing suitable target peptide bonds are available in the connecting segments. Sometimes, also, domains can be distinguished by scanning calorimetry because of differences in thermostability (7).

The ideal way of distinguishing domains is by direct observation of the three-dimensional structures obtained by X-ray diffraction. The number of known three-dimensional structures is still small relative to the number of available sequences, however, and most of the domains we are addressing here were uncovered on the basis of sequence analysis alone. The problem is complicated by the fact that amino acid sequences are eroded during evolution at a much faster rate than are three-dimensional structures (Figure 1). The number of known three-dimensional structures is growing rapidly, however, and in the future this will doubtless be the method of choice for establishing

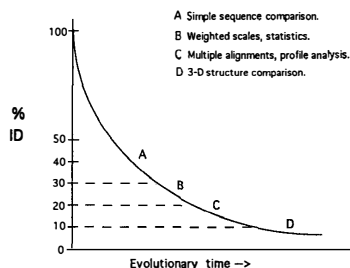


Figure 1 Divergence of two sequences from a common ancestor occurs as a negative exponential. If two (sufficiently long) sequences are more than 30% identical (A), simple visual inspection is enough to warrant a judgment about common ancestry. As the resemblance decays (B–D), increasingly sophisticated means are required to validate the relationship.

domain relationships. In the meantime, sequence searching methods have been extremely successful in identifying members of the major families.

Sequence Searching Programs

There are numerous computer programs available for searching an amino acid sequence against a database in an effort to find related sequences, but because the subject is relentlessly reviewed (8,9), only a brief comment is needed here. The most popular program currently in use is called Blast (10), having replaced FastA (11) as the most frequently used routine. If a sequence belongs to a known family, then a variety of pattern searches is available (12–14); in general, these will be more sensitive than the search of a single sequence. In any case, it always pays to conduct the operation iteratively, which is to say, the search should be reexecuted with the inclusion of whatever other sequences are initially uncovered.

Three-Dimensional Comparison Programs

A priori, we expect that there will be limits to the effectiveness of relationships based on amino acid sequences alone. Sequences change more or less stochastically, and eventually a point will be reached where not enough information remains to distinguish genuine homology from chance similarity (Figure 1). This is especially true in the case of smaller domains composed of fewer than a hundred residues, which includes many of the most common domains. Family memberships will in general be undercounted, and distant relationships may have to be established exclusively by three-dimensional structures.

A number of programs are available for searching three-dimensional struc-

tures against databases of other three-dimensional structures [recently reviewed by Orengo (15)]. Occasionally it is possible to find an unexpected relationship without a formal computer search. An interesting case in point has to do with the domain ("fold") now known as the cystine knot dimerization domain. The fold was originally observed in nerve growth factor (16). Subsequently, a structure appeared for the growth factor known as TGF-beta (17); it, too, was reported as a new fold. A sharp-eyed reader (18) saw a similarity between the two structures, however, and the two proteins have in fact turned out to be members of the same family (19). The structure of platelet-derived growth factor showed it to be related (20), and computer searching has now extended the family to include several connective tissue growth factors (21). A recent modelling study has revealed that the Norrie disease protein is also in this family, as are parts of the proteins mucin and von Willebrand factor (22). In all of these cases, the sequence resemblances are marginal at best.

Computer analysis is rapidly replacing the casual inspection of new three-dimensional structures. Actin is a protein so far known only in eukaryotic cells, and yet when its crystal structure was determined, to everyone's surprise, it was found to resemble the ATPase domain of heat shock 70, a ubiquitous chaperone protein (23). When the structure was scrutinized by careful computer searching, the family was extended to include not only hexokinase (23) but a number of other kinases as well (24). Moreover, phycocyanins turn out to have structures that are clearly similar to globins (25), and now the relationship has been extended to bacterial toxins (26). In another case, cold shock protein has been found to have the exact same fold as an oligonucleotide-binding protein (27), which in turn has the same structure as an oligosaccharide-binding protein (28, 29). Whether it will be possible to root today's multiplicity of domains back to their starter types on the basis of three-dimensional comparisons remains to be seen, but significant progress along these lines is already evident.

Classifying Domains

To a degree, how one classifies domains depends on one's notion of how they came into being in the first place. If, for example, all domains are descended from a single common ancestor, then a structure-based classification derived from a phylogenetic tree is an ideal representation. At the other extreme, if all domains have been concocted independently from some kind of random "big bang" (4, 30), then phylogenetic trees will be useless in this regard, and some kind of arbitrary classification will have to be imposed. A simple structure-based classification is widely used when three-dimensional information is available (31); assignments to classes can be made automatically by computer (15, 32). But three-dimensional data are often lacking, and schemes are needed

that follow functional, cytogeographic, or when appropriate, phylogenetic lines.

So far, no single means of classification has proven altogether satisfactory. Although there is some utility in discussing domains as being either extracellular or intracellular, or, in the case of eukaryotes, nuclear or otherwise compartmentalized, we now know of many exceptions to such efforts at cellular delimitization. Thus, the immunoglobulin domain was initially thought to be restricted to an extracellular existence, but it has now been found in intracellular proteins, too (33). Similarly, phylogenetic restrictions have turned out not to be as rigid as first believed (34); again, the immunoglobulin domain has been found in several bacterial proteins and may prove to be ubiquitous. Finally, in the past, domains have been roughly categorized into those that contain disulfide bonds and those that do not. Now we realize that there are several domain families in which some members have disulfides and some do not. Again, the immunoglobulin domain is a good example.

Nomenclature

The names of the individual domains present a problem; clearly, we are at a point where a systematic nomenclature needs to be imposed. Many domains have been named on the basis of the first protein in which they were found, whether or not that protein was a good representative. Others have been named arbitrarily by function, others by undecipherable acronyms that strain the imagination; several domains have more than one name. For the moment, the reader will have to bear with a certain amount of anagrammatic confusion. But the discovery of new domains is slowing, and the time has come for workers to agree on a simple comprehensive system (35).

EXONS, INTRONS, AND MODULES

There is a widespread but mistaken belief that all exons encode domains. This lore has been encouraged by the notion that primitive proteins were encoded by "minigenes" that were spliced together, the genome eventually maturing into a state where coding regions were separated by introns (36, 37). The popular "introns early" scenario (38) holds that the prokaryotic lineage has lost all (coding region) introns over the course of time as a secondary adaptation concordant with streamlining the genome or quickening replication times (39). The "introns late" doctrine, on the other hand, holds that the class of introns that interrupt protein coding regions are a latter-day invention, making their impact after the occasion of the major endosymbiotic events that led to the appearance of mitochondria in eukaryotes (40)

Structural Aspects

Many studies have been conducted that attempt to show that intron positions in contemporary protein coding regions correlate with rudimentary structures (41–45). Although some of the earliest of these studies were vague about what constituted basic structural elements, two recent efforts have tried to settle the point with a more rigorous approach in which observed intron distributions are measured against what would be expected by chance (46, 47). Both come down in favor of the null hypothesis: To wit, in ancient proteins there is little or no correlation of present-day intron positions with recognized elements of protein structure.

This is not to say that introns do not aid and abet the movement of defined domains about the genome. In many instances, particularly evident in animal proteins that have been fashioned within the last billion years, it is clear that introns have been greatly involved (48, 49). It may even prove to be the case that the enormous success of the metazoan radiation is the direct result of the introduction of introns into coding regions. All the same, it must be emphasized that most exons do not encode domains, and that many movable domains are in fact themselves interrupted by introns. In these latter cases, structural integrity demands that the full complement of exons be shuffled as a unit.

We must ask why the genes of some animal proteins fit so well with the concept of exon shuffling while others do not (Table 1). As a general rule, those that have well-delineated domains set off by introns have been fashioned more recently. Beyond that, the discordance in other instances may be attributable to two causes. First, some proteins may have been assembled by exon shuffling, but subsequent loss and gain of introns has obscured the relationship. Several of the proteins in Table 1 likely fall into this category. Second, some proteins may have been assembled in a modular fashion, but without the involvement of introns. Consider a typical zymogen protease such as trypsinogen: It has two well-defined and similarly folded domains, but none of its four introns occur at the obvious joining point (59). Clearly, a simple contiguous duplication gave rise to the structure, and at some later time the gene was invaded by introns.

Genetic Aspects of Domain Rearrangement

DNA replication is given to a variety of iterative errors, and chance duplications can range from a few base pairs to entire chromosomes and even genomes. Various kinds of recombination, homologous (legitimate) or illegitimate, occur as the result of random or nonrandom breakage and reunion of DNA. Homologous recombination is the result of mistaken mismatching of similar DNA sequences, and as a result, there is a tendency for duplication to beget more duplication (67). The more recombination occurs at a given locus,

Table 1 Correlations (or not) of animal protein domains with exons

Good correlation	Poor or no correlation
tissue plasminogen activator (50)	trypsinogen (59)
LDL-receptor (51)	serpins (60)
vonWillebrand factor (52)	complement C6 (61)
protein S (vitronectin) (53)	complement C3–C5 (62)
link protein (54)	α spectrin (63)
fibronectin (55)	preadipocyte EGF-like protein-1 (64)
factor XIII b chain (56)	notch (65)
homing receptor (57)	lin12 (66)
lipoprotein-associated coagulation inhibitor (58)	

the more opportunities there will be for mismatch. In compensation, intron sequences drift away much more rapidly than coding sequences, thereby diminishing opportunities for homologous recombination. On the other hand, introns often provide safe havens for transposable elements or highly repetitive sequences like the Alu family, and these in turn can become involved in mismatches that give rise to rearrangements (68). Additionally, there is a good deal of rearrangement by more obscure mechanisms that usually fall under the rubric “illegitimate recombination.”

There is an additional genetic constraint, as first pointed out by Patthy (69), imposed by the necessity for shuffled exons to have compatible introns so that neighboring domains will not find their coding to be out of register. Interestingly, the vast majority of introns flanking movable domains in animal proteins are of the phase I variety, the terminal codons of the separated exons being interrupted after the first nucleotide position (70).

THE PHYLOGENETIC DISTRIBUTION OF DOMAINS

Some Ancient Domains

This review is organized as though one could classify domains arbitrarily along phylogenetic and cytogeographical lines. In line with the author’s bias that the number of primordial domains will ultimately prove to be small, we begin with the ancient proteins, which have been defined as those proteins that are common to all or most creatures on the Earth today (34). It is presumed that such proteins were in existence before the last common ancestor, which is to say about two billion years ago, give or take a half a billion. In this regard, we must take note of the modular structure of ubiquitous mainstream enzymes that occur in the cytoplasm, mentioning, if only in passing, ligand-binding

Annu. Rev. Biochem. 1995.64:287-314. Downloaded from www.annualreviews.org by NORTH CAROLINA STATE UNIVERSITY on 01/16/13. For personal use only.

domains such as the nucleotide-binding fold (71) or the flavin-binding (72) or heme-binding (73) modules. Most of these domains are now locked into their positions in proteins at the core of metabolism, and the evidence of their ancient mobility is apparent only from three-dimensional comparisons (74), all sequence resemblance having been eroded long ago (although, as happens in the case of the nucleotide-binding fold, occasionally a weak sequence motif, or signature sequence, remains that can be diagnostic). In contrast, many non-catalytic domains still move around in various positions peripheral to catalytic domains, in prokaryotes and eukaryotes alike, as made clear by a number of recent examples (75).

During this discussion, we must be careful not to exaggerate the evolutionary power of shuffling modules about in proteins. Although the combinatorial advantage is significant, it is not unlimited, and it has sometimes proved more expedient to reinvent a feature than to depend on genetic shuffling. Nucleotide-binding folds, for example, appear to have been reinvented on several occasions (76). Furthermore, we must distinguish between a system in which only structural components are exchanged (4), on the one hand, and one in which already functional domains are swapped, on the other. In the latter case, one would expect only a small fraction of combinations ever to be useful, although there are some settings where the utilization of an already existing domain seems made to order. For example, the regulation of gene expression seems a perfect realm for combinatorics, a small number of DNA-binding domains being paired with many different ligands or metabolites.

In this regard, the modular nature of numerous DNA binding proteins from bacteria has been firmly established both by X-ray crystallographic studies and primary structure comparisons. In many bacterial repressor proteins, for example, there are DNA-binding domains associated with various ligand-binding domains (77). A compelling example is the catabolite gene activator protein (CAP), which binds to DNA in a sequence-dependent manner after it has been complexed with cAMP, and which has a distinctly separate cAMP-binding domain in tandem with a helix-turn-helix domain that is clearly homologous to many other DNA-binding proteins (78). Modular proteins are also at the heart of bacterial signaling systems, the transmitter-receiver basis of which is largely combinatorial (79). Prokaryotic enhancer proteins also have a eukaryote-like modularity (80).

In retrospect, it should not have been surprising that many of these domains are common to prokaryotes and eukaryotes. Thus, the helix-turn-helix structural motif, first observed in bacteria (81), has its counterpart in the eukaryotic homeodomain (82–84). The most graphic illustration of how structurally similar prokaryotic and eukaryotic DNA-binding domains can be is afforded by the virtual superimposability of the eukaryotic POU-specific domain (Pit-Oct-Unc) on the lambda repressor (85, 86)

Still, prokaryotes and eukaryotes may have some domains that are unique to themselves. Indeed, each may have major lineages that have domains that were fashioned along the way; a number of schemes have been put forth for how new structures might be conceived, including frame-shifts (87) and intron capture (88). Certainly fungi, plants, and animals, respectively, have many sequences that appear not to have counterparts elsewhere (see below). Either new domains have been concocted, or old ones have been selectively lost by some lineages, or, in many cases, relationships have been blurred by excessive sequence change, and the related structures have yet to be studied by three-dimensional approaches (Figure 1). For example, steroid-binding receptors, modular DNA-binding proteins with shuffled domains (89), are so far a strictly animal phenomenon (90, 91), although the possibility of their existence in fungi has been suggested (92).

Intracellular Modular Proteins Apparently Unique to Eukaryotes

One module with a characteristic sequence motif that has so far been found only in eukaryotes is the widespread WD-40 domain, first identified in beta-transducins from animals and cell-division-cycle proteins from yeast (93), but now known to occur in upwards of 50 different settings (94). This domain, which is only about 40 amino acids long, has a definable sequence pattern (X..GH..X..WD) that has been conserved over the billion years since fungi, plants, and animals last had a common ancestor; it has not been found in bacteria. The domain, which is usually found as a set of 4–8 repeats, is associated with a staggering number of functions and interactions with other proteins. Its proposed structure, modeled from similar sequences in parts of known structures (94), is a small five-stranded antiparallel sheet. Its evolutionary origin remains mysterious.

There are other apparently critical modules that are known to be common to fungi, plants, and animals, but which have not yet been traced to a prokaryotic origin. For example, the dimerization domain that came to be known as the leucine zipper (95) was first identified by sequence searching when it was found to be common to the carboxyl-terminal segment of the yeast GCN4 protein and vertebrate oncogenes like *jun* and *myc* (96); databases now contain scores of related sequences, but all of them are from eukaryotes.

Another set of widely distributed domains that is so far restricted to eukaryotes is involved in recognizing and binding phosphorylated proteins. The src homology region 2 (SH2) is known to bind phosphorylated tyrosines in proteins (97, 98). It frequently occurs in association with another well-defined domain, the src homology region 3 (SH3), the function of which is not yet clear (97). Interestingly, its three-dimensional structure (99) has now been

reported to resemble that of a photosystem protein from a cyanobacterium (100).

Eukaryotic cells also have a commonly shuffled domain called the PH domain (Pleckstrin Homology), which may be involved in the recognition and binding of phosphorylated serines and threonines (101). There is also a 42-amino acid repeat that occurs in a wide variety of fungal and animal cells that apparently participates in multiple interactions within the nucleus, including the maintenance of the nucleolus. In mammals it was first called plakoglobin (102); the *Drosophila* homologue corresponds to a mutant called *armadillo* (103), and the first observation in yeast involved a suppressor of temperature-sensitive mutations (104). It is also found in beta-catenin, an adhesive junction protein in animals (105). The number of repeats within any given setting may be as high as eight.

Finally, there is the 33-residue repeat known as the ankyrin repeat. Computer searching reveals that this domain is spread through a wide variety of fungal and animal proteins, and also in a few scattered bacteria (106). The distribution is unusual enough that the possibility of horizontal gene transfer has been raised. As in the case of other relatively short segments, the domains presumably gain stability by packing against each other, and as a result, the sequence motif never occurs as an isolated unit.

EXTRACELLULAR MOSAIC PROTEINS IN ANIMALS

It was the explosion of movable domains that occurred with the evolution of metazoans that offers us our richest demonstration of the construction of new proteins by shuffling modules. Here we have two great advantages. First, the intense research focus on animal proteins, combined with large-scale genome projects on humans, mice, fruit flies, and worms, has provided an enormous database. Second, and more importantly, the shuffling has occurred much more recently and is therefore more evident. Amino acid sequences alone are usually sufficient for following the dispersal of units. It appears, also, that the shuffling has been enormously more prevalent in animals and, indeed, has been part and parcel of the evolution of metazoan development, giving rise to the many unique proteins needed in cell-to-cell signaling and to the systems of defense and repair that are so essential to a multicellular existence (Figure 2).

Animal tissues are a marvel of structural complexity and intercellular communication. Many different cells exist in an extracellular matrix (ECM) that is itself an enormous assemblage of macromolecules, many of the components of which appear to be present in all animals but virtually nowhere else (113–115). The most abundant of these proteins is collagen, the highly repetitive structure of which was likely generated *de novo* as a part of animal emergence. The unexpected structural diversity of this protein, forming cables in some

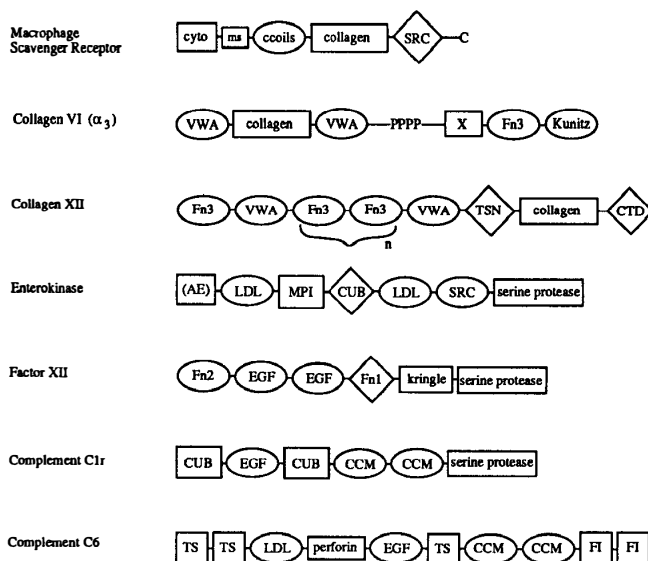


Figure 2 Some representative mosaic proteins. All the examples are from animal extracellular proteins (61, 107–112). Codes: CCM, complement control module; LDL, low-density lipoprotein receptor; Fn1, Fn2, and Fn3, fibronectin type I, II, and III repeats, respectively; EGF, epidermal growth factor-like; TS, thrombospondin type 1 repeat; TSN, thrombospondin amino-terminal domain; VWA, von Willebrand type A repeat; SRC, macrophage scavenger receptor domain; CUB, complement subcomponent; FI, factor I cryptic protease; MPI, metalloproteinase inhibitor; CTD, collagen carboxyl-terminal domain; AE, alternative exon; ms, membrane-spanner. Signal peptides are not shown; PPPP denotes proline-rich region. (Unfortunately, this figure was prepared before the recommendation on standardized abbreviations noted in Table 2.)

instances and sheets in others, is partly attributable to its being associated with a cadre of different noncollagen domains (Figure 2). In addition to collagen, there is a potpourri of syndecans and perlecan, laminins and mucins, all contributing to what used to be called “the ground substance,” the nonprotein portion of which has long been known to be rich in sulfated mucopolysaccharides. It turns out that many of the mobile modules are themselves attachment sites for glyco-sulfates (116–119), while many others bind to these heparinic materials.

Approximately 50 different domains have been identified that are shuffled about in various animal extracellular proteins (120, 121). Of these, a few are much more abundant than all the others combined. These include the so-called EGF-like domain and the immunoglobulin (Ig) and fibronectin type III (Fn3) domains (it should be noted that Ig and Fn3 are not restricted to animals or to

Table 2 Some modules occurring in extracellular proteins of animals that do not depend on disulfide bonds for their integrity

Code ^a	Module	Length (res.)	Reference ^b
FA58A (FA)	coagulation factor V/VIII, type A	300	122
VWFA (VA)	von Willebrand factor, type A	200	123
LAMG (LG)	laminin G-like (A-type module)	190	124
FA58C (FC)	coagulation factor V/VIII, type C	150	122
ClQ (CQ)	collagen/complement Clq	140	125
CADH (CA)	cadherin-like	130	126
IGSF (IG)	immunoglobulin	100	127
FN3 (F3)	fibronectin, type III	90	55
HEMOP (HX)	hemopexin-like	60	53
LDLY (LY)	"YWTP" repeat, LDL-receptor	50	51
LRP (LR)	leucine-rich (tollid)	30	12

^a An international workshop on extracellular protein modules was held in Angelholm, Sweden, September 24–28, 1994 (P Bork, S Forsen, A Pastore & M Sunnerhagen, organizers), during which the participants discussed and endorsed a standardized scheme for referring to some 30 different modules found in extracellular proteins. The common abbreviations use 3–5 characters, but a two-letter designation (in parentheses above) is recommended for use in cartoons. The complete list, which was mainly compiled by P Bork and A Bairoch, appears in Reference 128.

^b An effort has been made to cite the original observation of evolutionary mobility, but in some cases only the original sequence report is given, whether or not a module was described.

an extracellular existence). Other frequently shuffled modules include the calcium-dependent carbohydrate-recognition domain (C-lectins), complement control protein modules, cadherins, collagen segments, some domains originally observed in the LDL receptor, and about a dozen others (Tables 2 and 3). Most of these domains are found in each other's company, a mosaicism attributable to the genetic and structural compatability discussed above.

Some modules occur in remarkably long tandem repeats. The complement control protein module (also known as the beta-2 glycoprotein and sushi domain) occurs in six complement proteins, at least four complement regulatory proteins, and assorted noncomplement proteins as well. The complement control protein CR1 (not to be confused with C1r) is composed of 31 such domains strung together, and coagulation factor XIII b chain is wholly composed of 10 such repeats. Interestingly, a number of different proteins that contain this domain are found at the same locus on chromosome 1 in humans (140), suggesting that much of the duplication and shuffling is relatively recent. This chromosomal region may be likened to an active volcano from which further eruptions can be expected.

Many of these evolutionarily mobile domains are versatile in their binding properties. They form good packing units that bind to themselves as well as to other modules; many bind ligands. Their attachments to ligands may be loose and temporary or tight and permanent. They are seldom involved directly

Table 3 Some animal protein modules whose structure is dependent on disulfide bonds

Code ^a	Module	Length (res.)	Cysteines	Reference ^b
VWFB (VB)	von Willebrand factor, type B	30	8	124
SOMAB (SO)	somatomedin (vitronectin)	40	8	128
LDLRA (LA)	LDL receptor, type A	40	6	51
FN1 (F1)	fibronectin, type I	40	4	55
EGF (EG)	epidermal growth factor-like	40	6	129
FOLL1 (FS)	folliculin (ovomucoid)	50	6-10	130
PDOM (PD)	P domain (trefoil)	60	6	132
FN2 (F2)	fibronectin, type II	60	4	55
TSP1 (T1)	thrombospondin, type I	60	4-6	133
CCP (CP)	complement control protein (β_2 GP, sushi, SCR)	65	4	134
FIMAC (FM)	complement factor I, MAC proteins	70	8-10	135
CTCK (CK)	C-terminal cystine knot	90	6-11	18
KUNIT (KU)	Kunitz-type inhibitor	90	4-6	130
LINK (LK)	link protein/proteoglycan	100	4	54
SRCR (SR)	scavenger receptor	110	6	136
VWFC (VC)	von Willebrand factor, type C	110	10	124
CUB (CU)	complement subcomponent module	110	2-4	137
KRING (KR)	kringle	120	6	138
CLECT (CL)	C-type lectin	130	4-6	139

^a See footnote to Table 2.^b See footnote to Table 2.

in catalysis. Many are calcium dependent, consistent with the millimolar concentrations of calcium that occur in animal extracellular fluids. The majority depend on disulfide bonds for their structural integrity (Table 3), another advantage of an existence in the extracellular milieu.

Functionally, many are involved in cell recruitment, others in protein gathering or the assembling of other extracellular aggregates. The protein-protein interactions are often casual associations more than firm attachments. The specificity may be narrow or loose, the latter often reflecting a kind of cross-talk between systems or pathways. In animal bloods and fluids, mosaic proteins containing these various interacting domains are often fellow travelers rather than conjugal partners. In defense systems, these mosaics exhibit multiple interactions that keep a variety of substances in the vicinity of points of vulnerability such as vessel walls (141). Other aggregates seem to operate as a kind of macromolecular wolf pack, waiting in swarms for the inevitable microbial attack.

Many of these domains are genetically joined to proteases; others are themselves protease inhibitors (58, 131, 142). This is in keeping with the principal

mode of extracellular regulation involving limited proteolysis. The vertebrate blood coagulation and fibrinolysis schemes have been assembled primarily by shuffling appropriate modules into position at the amino-termini of serine proteases. On the other hand, many other mobile domains are involved in transmembrane signaling and constitute the extracellular portions of membrane receptors that are kinases or phosphatases, in keeping with intracellular regulation being mainly dependent on phosphorylation and dephosphorylation. Regulation emerges as a dominant role.

The EGF-Like Domain

Epidermal growth factor is a small (53 residues in human), stand-alone extracellular polypeptide stabilized by three disulfide bonds, which when bound to suitable cell surface receptors stimulates processes leading to mitosis and growth (143). The cloning of the unexpectedly large precursor of this protein (144, 145) revealed that the growth factor was one of a number of repetitive homologous units. Computer analysis further revealed that some of the homologues in the precursor were significantly related to doubly repeated segments in several proteins involved in blood coagulation (146).

Since that time, related sequences have been identified in more than 70 different vertebrate and invertebrate proteins (35), often occurring in long strings, and never in the cytoplasm, a natural consequence of their integrity depending on disulfide bonds. NMR studies have indicated that the structure is a compact bundle of five short beta strands (147). The family can be divided into two subclasses, one of which is ordinarily modified posttranslationally and binds calcium (148).

Recently, the domain has been found in a plant protein kinase (149). Even further afield, the domain has been found in the malaria-causing protozoan, *Plasmodium falciparum*, as a string of four units in one instance (150) and as a block of two in another (151). The sequence resemblance is sufficiently strong that there can be no doubt of homology, to the point where the question arises as to whether the parasite may have acquired the domain from some animal host (see below). Coincidentally, another protozoan, *Euplotes raikus*, makes a pheromone growth factor, the size of which is about the same as EGF and which contains six cysteines also (152), but the disulfide pairing is different (153) and NMR studies show it to be composed of several alpha helices (154).

The Immunoglobulin Domain

Some three decades ago it was proposed, on the basis of preliminary amino acid sequence data, that the heavy and light chains of immunoglobulins were not only homologous, but that these proteins had been assembled during the course of evolution by the extensive duplication of a primitive gene corresponding to a polypeptide chain of about 100 amino acids (155, 156). At the

time there was no indication that these structures would turn out to be among the most common of rudimentary protein modules, or even that similar structures would be involved in other aspects of the immune response. Today it is clear that the superfamily not only includes antibodies and T-cell receptors, but also hundreds of other proteins involved in animal cell-cell attachment and communication (127, 157). The original family relationship was based very much on the existence of paired cysteines for what was thought to be an essential disulfide bond. As it happens, many members of the family lack the disulfide, showing that the characteristic sandwich structure can be sustained without auxiliary stabilization (157).

The diversity of the proteins in which these domains are found is simply staggering. Many of them are membrane bound and are involved in all sorts of cell adhesion events. Many others serve as the extracellular portions of membrane-bound receptors, the cytoplasmic portions of which are often kinases or phosphatases (158). A remarkable number are also associated with the nervous system and are referred to by a bewildering array of acronyms (e.g. N-CAMs, MAG, L1, TAG-1) and fascinating cognates (e.g. contactin, fasciculin, neuroglian, amalgam) (159). Moreover, the same kinds of domains are components of intracellular proteins, such as the skeletal muscle C-protein (33) and a smooth muscle light chain kinase, as well as the long connection proteins titin (160), twitchin (161), and projectin (162).

For a long time, it was thought that the immunoglobulin domain was confined to animals. Indeed, extension of the family to invertebrates remained tentative for a considerable period (127). In 1989, however, an X-ray structure of a bacterial chaperonin protein revealed a structure that was exceedingly similar to the immunoglobulin fold (163). There is no apparent sequence resemblance between this protein and animal immunoglobulins, although the point is disputed. As we shall see, however, the bulk of the evidence favors common ancestry.

The Fibronectin Type III Domain

The fibronectin type III (Fn3) domain has a distinctive motif (164), the structure of which was predicted on the basis of a multiple alignment of many cytokines to be a seven-stranded beta sandwich similar to that of immunoglobulins (165). The prediction has subsequently been borne out by experiment (166, 167). Although the structure is very similar to the immunoglobulin fold, it is distinguished in that its strand arrangement is a three-on-four sandwich (EBA:GFCD) instead of the four-on-three (DEBA:GFC) found in the class typified by the immunoglobulin constant region. Fn3 units do not have disulfide bonds (Table 2).

The domain occurs in a wide variety of animal proteins, both extracellular and intracellular. It is especially common in the extracellular portions of

membrane-bound tyrosine kinase and tyrosine phosphatase receptors, where it is often found in the company of immunoglobulin domains. Similarly, the two of these domains together comprise the intracellular proteins titin, twitchin, and projectin (160–162). The domain has been estimated to occur in 2% of all animal proteins (168, 169); as we shall see, this may be a significant underestimate.

The domain has yet to be identified in plant or fungal proteins, but, remarkably, it has been readily identified by sequence analysis in a family of extracellular enzymes in bacteria, and the spectre of horizontal gene transfer has been raised to account for its existence in those enzymes (see below). The apparent absence from fungi and plants may be an artifact of the limits of sequence searching, however, even when a relatively sophisticated pattern search is employed. That many Fn3 domains were missed by sequence searching is apparent from the recent report of an Fn3 domain revealed by the X-ray structure of coagulation factor XIII (170). Similarly, the recently reported X-ray structure of beta galactosidase from *E. coli* showed that it has two domains of this type (171), and a previously reported cellulase from bacteria has one also (172). These structures will likely turn out to be ubiquitous, and it will be surprising indeed if three-dimensional studies of plant and fungal proteins do not turn them up eventually.

Although the Fn3 domain can be distinguished from the immunoglobulin domain both by sequence and three-dimensional criteria, the structural correspondence is so close that the two may be derived from a common ancestor. Indeed, a compelling study (173) based on a careful analysis of three-dimensional structures leads to the conclusion that all immunoglobulin and Fn3 domains arose from an ancestral form that most probably resembled the bacterial domains such as PapD.

Some Other Common Domains

A few other domains have been extensively shuffled about in animal extracellular proteins (Tables 2 and 3). Recently, there have been some excellent reviews focused on these modules, and only a few brief comments are made here about some of them. The nomenclature for many of these modules remains a problem. For example, the complement control protein modules were originally observed in a blood plasma protein of unknown function called the beta-2 glycoprotein and were initially referred to as B2GP repeats (134). They were subsequently found in enormous numbers in various complement proteins and have been rechristened as complement control modules. At about the same time, the discovery by Japanese workers of their presence in the hemolymph clotting system of the horseshoe crab led to them being dubbed “sushi domains” (174).

Lectin domains also have a dizzying nomenclature (175). The calcium-de-

pendent C-lectins, which occur either singly or in sets of two or three in widely diverse settings, have been reviewed recently (176), as have the collagen-associated collectins (177). The cadherins are another widely spread group that binds to a great diversity of substances in a calcium-dependent manner (178).

Two aspects about the large number of modern mosaic proteins need emphasizing: timeliness and structural compatability. That so many of them appear at about the same time in a phylogenetic sense and that so many of them are associated with each other is not a mere coincidence. Let me try to illustrate the connection with a few more comments about collagen domains and where they are found.

For a protein whose structure was thought to have been determined with finality a generation ago, collagen has turned out to exist in a bewildering variety of forms and in association with a cadre of other domains, including the Fn3 domain and various von Willebrand factor repeats (Figure 2). The Gly-Xxx-Pro repeat that typifies collagen is well known to form a tightly woven structure composed of three helices. Collagen occurs in two quite different settings, now called fibrillar and nonfibrillar. The fibrillar forms mature to very long stretches of the collagen motif flanked by globular procollagen segments that are removed by limited proteolysis. Remarkably, the fibrillar collagen gene in sponges has virtually the same exon arrangement as is found in mammals, indicating that the system has been in place since the very dawn of animals (179). The nonfibrillar forms are sometimes called "interrupted" and are found as mosaics in association with a variety of non-collagen domains (180–184).

Any noncollagen domains that find themselves situated immediately adjacent to collagenous segments, whether at their amino or carboxyl ends, must be able to accomodate a threefold symmetrical association. Casual examination reveals that this group includes the thrombospondin and von Willebrand C domains, both of which have cysteine spacings in common with the "pro-segment" of procollagen and are clearly homologous with it. Additionally, the headpieces of C1q obviously accomodate threefold packing ("pawnshop model"), as must the headpieces of the macrophage scavenger receptor family (185, 186). The 200-residue von Willebrand A domain, which lacks cysteine, is often found with collagenous segments on either side and must pack with a threefold axis. Moreover, members of the collectin family (177) have variably long collagenous segments with short coiled coil "necks" that separate them from carboxyl-terminal lectin domains; these, too, must obey the threefold rule. By extension, Fn3 units and kringles, which are also found in chains containing collagenous segments, must be able to pack similarly.

If we may continue with the theme of structural compatability, the number of reported mosaic proteins containing three-stranded alpha-helical coiled coils is also surprising. There was a time when only vertebrate fibrinogen was

thought to contain such a structure, but now the list includes various laminins (187, 188), thrombospondin (133), the collectins (177), the macrophage scavenger receptor (185) mentioned above, and a bevy of others. Scavenger receptor proteins were originally observed on the surfaces of macrophages (185). They are remarkable mosaics, combining compact globular domains with extended structures once thought to typify fibrous proteins (Figure 2). The MSCR module has now been found in several dozen different secreted or membrane-associated proteins, mostly involved in host defense (186).

Having just mentioned the three-stranded coiled coils that occur in fibrinogen, I should add that these are associated with large (250 residues) carboxyl-terminal domains previously referred to as FReDs, for fibrinogen-related domains (189). These are found in a number of extracellular proteins in both invertebrate and vertebrate animals. Most of these occurrences have been reviewed recently (189), but additional reports continue to pour in (190, 191), the most interesting of which is for the protein ficolin, which is a hybrid of collagen and fibrinogen (192).

Fibrinogen, of course, is the precursor of the fibrin blood clot, and I would be remiss if I did not say a few words about the interaction of fibrin with other substances by way of selected domains. Fibronectin (55) received its name because of this interaction, which is now mostly attributable to its FnI domains. But another moveable domain has been implicated for some other proteins. Thus, kringles were first identified in prothrombin and subsequently in several other proteins involved in blood clotting and fibrinolysis. In several of these proteins, the ability to bind to fibrin has been attributed to one or more of these domains. The most astonishing occurrence involves apolipoprotein a, which is a plasminogen homologue that contains 30 or more kringles at its amino terminus. In humans, apolipoprotein a isoforms with excessive numbers of kringles may correlate with thrombotic events (193). Beyond that, there have been some interesting attempts to follow the diaspora of kringle structures by sequence alignment and computer analysis (194, 195).

Most of the domains mentioned in the preceding paragraphs share a number of features. They all appear to exist in both vertebrate and invertebrate animals, and most of them appear to be able to coexist with each other. They dominate certain physiological systems peculiar to animals, including blood clotting and the immune system. Clearly, the advent of metazoans was a boom time for mixing and matching domains.

MODULES UNIQUE TO PLANT OR FUNGAL PROTEINS

Plants, animals, and fungi last shared a common ancestor about a billion years ago, and, as we noted earlier, they share many modular proteins that are involved in gene expression and intracellular regulation. But the three king-

doms have adapted quite different strategies for many of their extracellular involvements. Accordingly, it is perhaps not so surprising that most of the commonly shuffled domains that occur in extracellular proteins in animals have not yet been identified in plants or fungi. An interesting exception was the discovery of a somatomedin domain and four obvious hemopexin-type sequence repeats in the cytosolic plant protein PA2, a major storage protein component of pea seeds, leading to the suggestion that these two domains existed in the common ancestor of plants and animals (196).

Most of the other common domains (Tables 2 and 3), including anything resembling an immunoglobulin domain, remain among the missing. This may be partly attributable to the fact that the number of reported plant and fungal sequences is still relatively small, or it may be that the sequences have changed to the extent where they will only be recognized by three-dimensional studies, as has proven to be the case for some bacterial homologues. There is no indication that the somatomedin and hemopexin domains are changing any less rapidly than the others, however.

On the other hand, highly repetitive proteins that appear to be unique to plants are well known (197), and it seems reasonable that through one route or another some new modules have evolved during the time since plants last shared ancestry with animals. As for the fungi, the size and breadth of the ongoing yeast megasequencing projects tends to negate the notion that not enough data have been accumulated to find the missing domains. Rather, the surprise so far is that anywhere from a third to a half of all open reading frames (orfs) appear to be unidentifiable and unique to fungi (198). Again, this may reflect enormous sequence change that exceeds the limits of sequence searching. It is curious, though, what a large fraction of the yeast unidentified orfs resemble other unidentified open frames in yeast (198). These findings tend to confirm an earlier observation of an extensive family of proteins typified by a 34-residue repeat that also appeared limited to fungi (199).

EVOLUTIONARILY MOBILE DOMAINS IN BACTERIA

We have already commented that numerous ubiquitous proteins, principally those involved in mainstream metabolism, were constructed in a modular fashion in ancient times, as shown by obviously rearranged domains revealed by X-ray crystallography. That this mode of invention has continued throughout the evolutionary history of prokaryotes, including very recent happenings, is evidenced by the many examples of module shuffling observed on the basis of amino acid sequence similarity alone.

Thus, there are numerous instances of repetitive or transposed parts of proteins that occur in different settings. For example, the ribosomal protein S1, as originally sequenced in *E. coli*, was found to be composed, in its entirety,

of six homologous segments about 90 residues long (200). This was an unexpected situation, because functionally the protein was thought to be composed of an mRNA-binding domain, on the one hand, and a ribosomal-binding domain, on the other. Somehow the same primitive structure had adapted to both of these functions (both involve RNA binding, of course). Given the nature of the repeat, it was not surprising, later, when the enzyme polynucleotide phosphorylase was found to have a carboxyl-terminal domain that is clearly homologous with the S1 repeats (201). A simple phylogenetic tree suggests that the enzyme acquired it from the ribosomal protein and not the other way around, in that the enzyme domain is actually more similar to two of the S1 repeats than those repeats are to the others.

Another example of domain swapping uncovered by sequence analysis involves the *E. coli* DNA repair protein UvrA. This large protein (940 residues) is composed of a series of repeated segments that are clearly homologous with a family of periplasmic transport proteins (202). More recent computer surveys have expanded the superfamily greatly (203).

The recency of some of these exchange events can be gauged by examining how similar repeated segments are in different genomes. The problem is complicated in that many of the recently shuffled domains also seem to be involved in horizontal gene transfers. In many soil bacteria, carbohydrate-binding domains are found at widely differing locations in a diversity of extracellular glycosidases (204–206). Some of these domains have actually been observed in an occasional eukaryote (207). Furthermore, fibronectin type III (Fn3) domains are strewn about many of these same enzymes (168, 208), and they are simply too similar in sequence to animal Fn3 units to have been passed in a conventional manner from a common ancestor (168, 209). These proteins bear all the earmarks of “exon shuffling without introns.”

A number of prokaryotic structural proteins are also built up in an obvious domain fashion, including wall-attachment proteins in various bacteria. Some of these are long, multidomained proteins that contain coiled coil segments (210, 211). Others are highly repetitive and exhibit sequence motifs observed in other microbial proteins (212).

A fascinating example of module shuffling in the fashioning of bacterial mosaic proteins is offered by a set of membrane proteins that bind animal blood plasma proteins. A number of these are known that bind specific animal proteins with great specificity, including binders for fibrinogen, fibronectin, immunoglobulins, and plasma albumin. The staphylococcal protein A that binds to IgG, for example, is well known to most molecular biologists because of its great experimental utility. Less well known is that there are streptococcal proteins that not only bind to IgG, but also bind to plasma albumin. Protein G, for example, binds IgG with its amino-terminal domain and albumin with its carboxyl-domain (213). Moreover, certain strains of *Peptostreptococcus* are

able to bind immunoglobulin light chains with a protein denoted protein L (214). Other strains of this bacterium have a mosaic protein (PAB) that is clearly the result of very recent shuffling together of domains from proteins G and L and that binds albumin (215). The mechanisms for moving these domains about is presumed to involve homologous or nonhomologous recombination, or perhaps a conjugative plasmid (215).

Many other examples of rearranged modules in bacterial proteins could be cited, including the components of various nitrogen fixation pathways (216) or the prokaryotic phosphotransferase system (217), but the point should have been made by now that domain shuffling has been a persistent activity among bacteria.

SOME UNEXPECTED OCCURRENCES

Earlier, it was mentioned that tandem arrays of EGF-like domains are found in the malarial parasite, *Plasmodium falciparum* (150, 151). The question arises, Does this indicate the existence of this domain in common ancestors of protists and the later diverging eukaryotes? If that were the case, we would expect to find similar arrays in plants and fungi, where they are rarely observed. Did the parasite acquire the gene for such a domain from some animal host? Or is the remarkable sequence resemblance the result of sequence convergence? Arguments have been made against such unlikely occurrences (218).

Several other evolutionarily mobile domains have been found in this same protozoan, including a well-defined thrombospondin domain with a remarkable resemblance to one that occurs in the complement protein properdin (219, 220) and a domain found in coagulation factor VIII (120). Again, some kind of horizontal acquisition may have occurred. It should be mentioned also that the trypanosome *T. cruzi* has been reported to have a repeat similar to one that occurs in the LDL receptor, and also an Fn3 module (221).

An EGF-like domain has also been identified in vaccinia virus (222). Interestingly, vaccinia also contains another of the commonly swapped domains that appears in Table 3, the complement control protein module (223). Pox-viruses, of which vaccinia is one, are well known to acquire genes and pieces of genes from their mammalian hosts, and several other structures cited in this review have been found in their genomes also (75, 106)

HOW MANY DOMAINS?

Each month a popular biochemical journal contains a section devoted to claims and counter-claims of new domainal relationships based on sequence resemblances (224–232). Where will it end? Recent estimates about the total number of protein structures we can expect, based on the number of unique folds

already in the structure data bank, are in the range of 500–1000 (233, 234). There are a few hundred folds known at present (15), and many new reported structures are being assigned to already known families. Furthermore, as in the case of sequence analyses, it is now being recognized that what were thought to be separate groups may indeed share common ancestry (235). The question is not how many unique groups can one point to today, but rather how many primordial groups were there in the beginning? As an example, immunoglobulin domains can be distinguished from Fn3 domains, but the very close structural resemblance between the two families indicates strongly that they share common ancestry (173).

What is the minimum number of starter types that can account for today's inventory of domains? The number may be much smaller than the estimates based on current inventories, perhaps less than 20 and maybe as few as 10. These numbers emerge from a consideration of fundamental domain structures, which largely reduces to about 20 major types (31). These classes can be reduced further to two or three all-alpha, a few all-beta, and a small number of alpha-beta structures.

The dilemma we face is distinguishing structural convergence from evolutionary divergence (218). Are there so many parallel alpha-beta barrels because this is a fundamentally sturdy structure? Or, is it because it evolved once, was found to be sturdy and functional, and has been propagated by duplication and modular exchange ever since (236)? In the absence of data to the contrary, the latter explanation seems more reasonable (237).

Furthermore, some contemporary folds must have been fashioned since those protean times. In particular, disulfide-stabilized domains, which are so well represented in the extracellular mosaic proteins of animals (Table 3), probably could not have been concocted until after the Earth's atmosphere became oxidizing. Witness the absence of most of these domains from prokaryotes and cytoplasmic environments in general. Almost certainly, also, many of them will prove to have descended from common ancestors.

SUMMARY AND CONCLUSIONS

Domainal shuffling has occurred throughout the course of protein evolution. There are certain natural realms where the combinatorial advantage of mixing and matching domains can be very advantageous, but there are limits to its usefulness. Because amino acid sequences relentlessly diverge, the history of genetic rearrangements affecting the structure of ancient proteins is often revealed only by three-dimensional studies. In more recently evolved proteins, shuffled modules are often readily apparent on the basis of sequence considerations alone. In many modern animal proteins, it is clear that the rearrangement process has been greatly accelerated by the participation of introns. This

is not the same as saying that all exons encode domains, however. Certainly the majority of exons in eukaryotic genes do not. Moreover, many evolutionarily mobile domains, while genetically flanked by introns, also contain internal introns that do not participate in the rearranging. Several examples of recently rearranged modules in bacterial proteins are now known where domain shuffling obviously occurred without introns, and there is no reason to think the same prosaic processes were not at work in ancient times.

ACKNOWLEDGEMENTS

I thank Peer Bork for numerous discussions about protein modules and for sending me materials in advance of their publication. I am grateful to Karen Anderson and Chad Manning for help in the preparation of the manuscript.

Any *Annual Review* chapter, as well as any article cited in an *Annual Review* chapter, may be purchased from the Annual Reviews Preprints and Reprints service.
1-800-347-8007; 415-259-5017; email: arpr@class.org

Literature Cited

1. Wetlauffer DB. 1973. *Proc. Natl. Acad. Sci. USA* 70:697-701
2. Murzin AG. 1994. *Curr. Opin. Struct. Biol.* 4:441-49
3. Sommer SS, Cohen JE. 1980. *J. Mol. Evol.* 15:37-57
4. Dorit RL, Schoenbach L, Gilbert W. 1990. *Science* 250:1377-82
5. Doolittle RF. 1981. *Science* 214:149-59
6. Patthy L. 1991. *Curr. Opin. Struct. Biol.* 1:351-61
7. Donavan JW, Mihalyi E. 1974. *Proc. Natl. Acad. Sci. USA* 71:4125-28
8. Doolittle RF. 1994. *Curr. Opin. Biotechnol.* 5:24-28
9. Altschul SF, Boguski MS, Gish W, Wootton JC. 1994. *Nat. Genet.* 6:119-29
10. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. *J. Mol. Biol.* 215:403-10
11. Pearson WR, Lipman DJ. 1988. *Science* 85:2444-48
12. Patthy L. 1987. *J. Mol. Biol.* 198:567-77
13. Gribskov M, McLachlan AD, Eisenberg D. 1987. *Proc. Natl. Acad. Sci. USA* 84:4355-58
14. Rohde K, Bork P. 1993. *CABIOS* 9:183-89
15. Orengo C. 1994. *Curr. Opin. Struct. Biol.* 4:429-40
16. McDonald NQ, Lapatto R, Murray-Rust J, Gunning J, Wlodawer A, Blundell TL. 1991. *Nature* 354:411-14
17. Daopin S, Piez KA, Ogawa Y, Davies DR. 1992. *Science* 257:369-73
18. Swindells M. 1992. *Science* 258:1160-61
19. Daopin S, Cohen GH, Davies D. 1992. *Science* 258:1161-62
20. Oefner C, D'Arcy A, Winkler FK, Eggiman B, Hosang M. 1992. *EMBO J.* 11:3921-26
21. Bork P. 1993. *FEBS Lett.* 327:125-30
22. Meitinger T, Meindl A, Bork P, Rost B, Sander C, et al. 1993. *Nat. Genet.* 5:376-80
23. Flaherty KM, McKay DB, Kabsch W, Holmes KC. 1991. *Proc. Natl. Acad. Sci. USA* 88:5401-45
24. Bork P, Sander C, Valencia A. 1992. *Proc. Natl. Acad. Sci. USA* 82:7290-94
25. Pastore A, Lesk AM. 1990. *Proteins: Struct. Funct. Genet.* 8:133-55
26. Holm L, Sander C. 1991. *FEBS Lett.* 315:301-6
27. Schnuchel A, Wilschek R, Czisch M, Herrler M, Willmsky G, et al. 1993. *Nature* 364:169-71
28. Schindelin H, Marahiel MA, Heinemann U. 1993. *Nature* 364:164-68
29. Murzin AG. 1993. *EMBO J.* 12:861-67
30. White SK. 1994. *Annu. Rev. Biophys. Biomol. Struct.* 23:407-39
31. Richardson JS. 1981. *Adv. Protein Chem.* 34:167-339
32. Holm L, Ouzounis C, Sander C, Tuparev G, Vriend G. 1992. *Protein Sci.* 1:1691-98
33. Einheber S, Fischman DA. 1990. *Proc. Natl. Acad. Sci. USA* 87:2157-61

34. Doolittle RF. 1992. *Protein Sci.* 1:191-200
35. Campbell ID, Bork P. 1993. *Curr. Opin. Struct. Biol.* 3:385-92
36. Seidel HM, Pompliano DL, Knowles JR. 1992. *Science* 257:1489-90
37. Dorit RL, Gilbert W. 1991. *Curr. Opin. Struct. Biol.* 1:973-77
38. Gilbert W, Glynias M. 1993. *Gene* 135:137-44
39. Doolittle WF. 1978. *Nature* 272:581-82
40. Palmer JD, Logsdon JM. 1991. *Curr. Opin. Gen. Dev.* 1:470-77
41. Lonberg N, Gilbert W. 1985. *Cell* 40:81-90
42. Michaelson AM, Blake CCF, Evans ST, Orkin SH. 1985. *Proc. Natl. Acad. Sci. USA* 82:6965-69
43. Straus D, Gilbert W. 1985. *Mol. Cell. Biol.* 5:3497-506
44. Burke J, Hwang P, Anderson L, Lebo R, Gorin F, Fletcher R. 1987. *Proteins: Struct. Funct. Genet.* 2:177-87
45. Gilbert W, Marchionni M, McKnight G. 1986. *Cell* 46:151-54
46. Weber K, Kabsch W. 1994. *EMBO J.* 13:1280-86
47. Stolz J, Spencer DF, Zuker M, Logsdon JM Jr, Doolittle WF. 1994. *Science* 265:202-7
48. Gilbert W. 1978. *Nature* 271:501
49. Doolittle RF. 1985. *Trends Biochem. Sci.* 10:233-37
50. Ny T, Elgh F, Lund B. 1984. *Biochemistry* 81:5355-59
51. Sudhof TC, Russell DW, Goldstein JL, Brown MS, Sanchez-Pescador R, Bell GI. 1985. *Science* 228:893-95
52. Mancuso DJ, Tuley EA, Westfield LA, Worrall NK, Shelton-Inloes BB, et al. 1989. *J. Biol. Chem.* 264:19514-27
53. Jenne D, Stanley KK. 1987. *Biochemistry* 26:6735-42
54. Kiss I, Deak F, Mestric S, Delius H, Soos J, et al. 1987. *Biochemistry* 84:6399-403
55. Patel RS, Odermatt E, Schwarzbauer JE, Hynes RO. 1987. *EMBO J.* 6:2565-72
56. Bottenus RE, Ichinose A, Davie EW. 1990. *Biochemistry* 29:11195-209
57. Dowbenko DJ, Diep A, Taylor BA, Lusis AJ, Lasky LA. 1991. *Genomics* 9:270-77
58. van der Logt CPE, Reitsma PH, Bertina RM. 1991. *Biochemistry* 30:1571-77
59. Craik CS, Choo Q-L, Swift GH, Quinto C, Rutter WJ. 1984. *J. Biol. Chem.* 259:14255-64
60. Wright HT. 1993. *J. Mol. Evol.* 36:136-43
61. Hobart MJ, Fernie B, DiScipio RG. 1993. *Biochemistry* 32:6198-205
62. Vik DP, Amiguet P, Moffat GI, Fey M, Amiguet-Barras F, et al. 1991. *Biochemistry* 30:1080-85
63. Kotula L, Laury-Kleintop LD, Showe L, Sahr K, Linnenbach AJ, et al. 1991. *Genomics* 9:131-40
64. Smas CM, Green D, Sul HS. 1994. *Biochemistry* 33:9257-65
65. Wharton KA, Johansen KM, Xu T, Aravanis-Tsakonas S. 1985. *Cell* 43:567-81
66. Greenwald I. 1985. *Cell* 43:583-90
67. Doolittle RF. 1979. Protein evolution. In *The Proteins*, ed. H Neurath, RL Hill, Vol. 4:1-118. New York: Academic. 2nd ed.
68. Stoppa-Lyonnet D, Carter PE, Meo T, Tosi M. 1990. *Proc. Natl. Acad. Sci. USA* 87:1551-55
69. Pathy L. 1987. *FEBS Lett.* 214:1-7
70. Pathy L. 1994. *Curr. Opin. Struct. Biol.* 4:383-92
71. Rossmann MG, Moras D, Olsen KW. 1974. *Nature* 250:194-99
72. Correll CC, Ludwig ML, Bruns CM, Karplus PA. 1993. *Prot. Sci.* 2:2112-33
73. Vasudevan SG, Armarego WLF, Shaw DC, Lilley PE, Dixon NE, Poole RK. 1991. *Mol. Gen. Genet.* 226:49-58
74. Smith GCM, Tew DG, Wolf CR. 1994. *Proc. Natl. Acad. Sci. USA* 91:8710-14
75. Bork P, Doolittle RF. 1994. *J. Mol. Biol.* 236:1277-82
76. Schulz GE. 1992. *Curr. Opin. Struct. Biol.* 2:61-67
77. Bustos SA, Schleif RF. 1993. *Proc. Natl. Acad. Sci. USA* 90:5638-42
78. Steitz TA, Ohlendorf DH, McKay DB, Anderson WF, Matthews BW. 1982. *Proc. Natl. Acad. Sci. USA* 79:3097-100
79. Parkinson JS, Kofoid EC. 1992. *Annu. Rev. Genet.* 26:71-112
80. North AK, Klose KE, Stedman KM, Kustu S. 1993. *J. Bacteriol.* 175:4257-73
81. Sauer RT, Youcum RR, Doolittle RF, Lewis M, Pabo CO. 1982. *Nature* 298:447-51
82. Laughon A, Scott MP. 1984. *Nature* 310:25-31
83. Quian YQ, Billeter M, Otting G, Muller M, Gehring, Wuthrich K. 1989. *Cell* 59:573-80
84. Affolter M, Percival-Smith A, Muller M, Billeter M, Qian YQ, et al. 1991. *Cell* 64:879-80
85. Assa-Munt N, Mortishire-Smith RJ, Aurora R, Herr W, Wright P. 1993. *Cell* 73:193-205
86. Holm L, Sander C. 1993. *J. Mol. Biol.* 233:123-38
87. Keese PK, Gibbs A. 1992. *Proc. Natl. Acad. Sci. USA* 89:9489-93

88. Golding GB, Tsao N, Pearlman RE. 1994. *Proc. Natl. Acad. Sci. USA* 91: 7506-09
89. Laudet V, Hanni C, Coll J, Catzeflis F, Stehelin D. 1992. *EMBO J.* 11:1003-13
90. Evans RM. 1988. *Science* 240:889-95
91. Green S, Chambon P. 1988. *Trends Genet.* 4:309-14
92. Moore DD. 1990. *New Biol.* 2:100-5
93. Fong HKW, Hurley JB, Hopkins RS, Miake-Lye R, Johnson MS, et al. 1986. *Proc. Natl. Acad. Sci. USA* 83:2162-2166
94. Neer EJ, Schmidt CJ, Manbudripad R, Smith TF. 1994. *Nature* 371:297-300
95. Landschulz WH, Johnson PF, McKnight SL. 1988. *Science* 240:1759-64
96. Vogt PK, Bos TJ, Doolittle RF. 1987. *Proc. Natl. Acad. Sci. USA* 84:3316-19
97. Koch CA, Anderson D, Moran MF, Ellis C, Pawson T. 1991. *Science* 252:668-74
98. Waksman G, Kominos D, Robertson SC, Pant N, Baltimore D, et al. 1992. *Nature* 358:646-53
99. Fry MJ, Panayotou G, Booker GW, Waterfield MD. 1993. *Protein Sci.* 2: 1785-97
100. Falzone CJ, Kao YH, Zhao J, Bryant DA, Lecomte JTJ. 1994. *Biochemistry* 33:6052-62
101. Musacchio A, Gibson T, Rice P, Thompson J, Saraste M. 1993. *Trends Biochem. Sci.* 18:343-48
102. Franke WW, Goldschmidt MD, Zimbelmann R, Mueller HM, Schiller DL, Cowin P. 1989. *Proc. Natl. Acad. Sci. USA* 86:4027-31
103. Peifer M, Wieschaus E. 1990. *Cell* 63: 1167-78
104. Yano R, Oakes ML, Tabb MM, Nomura M. 1994. *Proc. Natl. Acad. Sci. USA* 91:6880-84
105. Peifer M, McCrea PD, Green KJ, Wieschaus E, Gumbiner BM. 1992. *J. Cell. Biol.* 118:681-91
106. Bork P. 1993. *Proteins: Struct. Funct. Genet.* 17:363-74
107. Kodama T, Freeman M, Rohrer L, Zabrecky J, Matsudaira P, Krieger M. 1990. *Nature* 343:531-35
108. Bonaldo P, Colombatti A. 1989. *J. Biol. Chem.* 264:20235-39
109. Yamagata M, Yamada KM, Yamada SS, Shinomura T, Tanaka H, et al. 1991. *J. Cell Biol.* 115:209-21
110. Kitamoto Y, Yuan X, Wu Q, McCourt DW, Sadler JE. 1994. *Proc. Natl. Acad. Sci. USA* 91:7588-92
111. McMullen BA, Fujikawa K. 1985. *J. Biol. Chem.* 260:5328-41
112. Leytus SP, Kurachi K, Sakariassen KS, Davie EW. 1986. *Biochemistry* 25: 4855-63
113. Schmid V, Bally A, Beck K, Haller M, Schlage WK, Weber C. 1991. *Hydrobiologia* 216/217:3-10
114. Irwin DH. 1991. *TREE* 6:131-34
115. Miklos GLG, Campbell HD. 1992. *Curr. Opin. Genet. Dev.* 2:902-6
116. Goldstein LA, Zhou DFH, Picker LJ, Minty CN, Bargatze RF, et al. 1989. *Cell* 56:1063-72
117. Fisher LW, Termine JD, Young MF. 1989. *J. Biol. Chem.* 264:4571-76
118. Bernfield M, Kokenyesi R, Kato M, Hinkes MT, Spring J, et al. 1992. *Annu. Rev. Cell Biol.* 8:365-93
119. Hinkes MT, Goldberger OA, Neumann PE, Kokenyesi R, Bernfield M. 1993. *J. Biol. Chem.* 268:11440-48
120. Bork P. 1991. *FEBS Lett.* 286:47-54
121. Bork P. 1992. *Curr. Opin. Struct. Biol.* 2:413-21
122. Vehar GA, Keyt B, Eaton D, Rodriguez H, O'Brien DP, et al. 1984. *Nature* 312:337-42
123. Shelton-Inloes BB, Titani K, Sadler JE. 1986. *Biochemistry* 25:3164-71
124. Beck K, Hunter I, Engel J. 1990. *FASEB J.* 4:148-60
125. Reid KBM. 1974. *Biochem. J.* 141:189-203
126. Shirayoshi Y, Hatta K, Hosoda M, Tsunasawa S, Saitoyama F, Takeichi M. 1986. *EMBO J.* 5:2485-88
127. Williams A. 1987. *Immunol. Today* 8: 298-303
128. Bork P, Bairoch A. 1995. *Trends Biochem. Sci.* In Press
129. Jenne D, Stanley KK. 1985. *EMBO J.* 4:3153-57
130. Doolittle RF, Feng DF, Johnson MS. 1984. *Nature* 307:558-60
131. Laskowski M Jr, Kato I. 1980. *Annu. Rev. Biochem.* 49:593-626
132. Thim L. 1989. *FEBS Lett.* 250:85-90
133. Lawler J, Hynes RO. 1986. *J. Cell. Biol.* 103:1635-48
134. Lozier J, Takahashi N, Putnam FW. 1984. *Proc. Natl. Acad. Sci. USA* 81: 3640-44
135. Larson RS, Corbi AL, Berman L, Springer T. 1989. *J. Cell. Biol.* 108:703-12
136. Freeman M, Ashkenas J, Rees DJG, Kingsley DM, Copeland NG, et al. 1990. *Proc. Natl. Acad. Sci. USA* 87: 8810-14
137. Bork P, Beckmann G. 1993. *J. Mol. Biol.* 231:539-45
138. Sottrup-Jensen L, Claeys H, Zajdel M, Petersen TE, Magnusson S. 1978. *Progr. Chem. Fibrinolysis Thrombolysis* 3:191-209
139. Drickamer K. 1988. *J. Biol. Chem.* 263: 9557-60
140. Carroll MC, Alicot EM, Katzman PJ,

- Klickstein LB, Smith JA, Fearon DT. 1988. *J. Exp. Med.* 167:1271-80
141. Preissner KT, Jenne D. 1991. *Thrombosis Haemost.* 66:123-32
142. Bolander ME, Young ME, Fisher LW, Yamada Y, Termine FD. 1988. *Proc. Natl. Acad. Sci. USA* 85:2919-23
143. Carpenter G, Cohen S. 1990. *J. Biol. Chem.* 265:7709-12
144. Gray A, Dull TJ, Ullrich A. 1983. *Nature* 303:722-25
145. Scott J, Urdea M, Quiroga M, Sanchez-Pescador R, Fong N, et al. 1983. *Science* 221:236-40
146. Doolittle RF, Feng DF, Johnson MS. 1984. *Nature* 307:558-60
147. Montelione G, Wuthrich K, Burgess AW, Nice EC, Wagner G, et al. 1992. *Biochemistry* 31:236-49
148. Rees DJ, Jones IM, Handford PA, Esnouf MP, Smith KJ, Brownlee GG. 1988. *EMBO J.* 7:2053-61
149. Kohorn BD, Lane L, Smith TA. 1992. *Proc. Natl. Acad. Sci. USA* 89:10989-92
150. Kaslow DC, Quakyi IA, Syn C, Raum MG, Keister DB, et al. 1988. *Nature* 333:74-76
151. Blackman MJ, Ling IT, Nicholls SC, Holder AA. 1991. *Mol. Biochem. Parasitol.* 49:29-34
152. Raffioni S, Miceli C, Vallesi A, Chowdhury SK, Chait BT, et al. 1992. *Proc. Natl. Acad. Sci. USA* 89:2071-75
153. Stewart AE, Raffioni S, Chaudhary T, Chait BT, Luporini P, Bradshaw RA. 1992. *Protein Sci.* 1:777-85
154. Brown LR, Mronga S, Bradshaw RA, Ortenzi C, Luporini P, Wuthrich K. 1993. *J. Mol. Biol.* 231:800-16
155. Singer SJ, Doolittle RF. 1966. *Science* 153:13-25
156. Hill RL, Delaney R, Fellows RE Jr, Lebowitz HE. 1966. *Proc. Natl. Acad. Sci. USA* 56:1762-69
157. Williams AF, Barclay AN. 1988. *Annu. Rev. Immunol.* 6:381-405
158. Yoshihara Y, Oka S, Ikeda J, Mori K. 1991. *Neurosci. Res.* 10:83-105
159. Krueger NX, Saito H. 1992. *Proc. Natl. Acad. Sci. USA* 89:7417-21
160. Labeit S, Barlow DP, Gautel M, Gibson TJ, Holt J, et al. 1990. *Nature* 345:273-76
161. Ayme-Southgate A, Vigoreaux J, Benlian G, Pardue ML. 1991. *Proc. Natl. Acad. Sci. USA* 88:7973-77
162. Higgins DG, Labeit S, Gautel M, Gibson TJ. 1993. *J. Mol. Evol.* 38:395-404
163. Holmgren A, Branden C-I. 1989. *Nature* 342:248-51
164. Pathy L. 1990. *Cell* 61:13-14
165. Bazan JF. 1990. *Proc. Natl. Acad. Sci. USA* 87:6934-69
166. de Vos AM, Ultsch M, Kossiakoff AA. 1992. *Science* 255:306-12
167. Leahy DJ, Hendrickson WA, Aukhil I, Erickson HP. 1992. *Science* 258:987-91
168. Bork P, Doolittle RF. 1992. *Proc. Natl. Acad. Sci. USA* 89:8990-94
169. Bork P, Doolittle RF. 1993. *Protein Sci.* 2:1185-87
170. Yee VC, Pedersen LC, Le Trong I, Bishop PD, Stenkamp RE, Teller DC. 1994. *Proc. Natl. Acad. Sci. USA* 91:7296-300
171. Jacobson RH, Zhang X-J, DuBose RF, Matthews BW. 1994. *Nature* 369:761-66
172. Juy M, Amit AG, Alzari PM, Poljak RJ, Claeysens M, et al. 1992. *Nature* 357:89-91
173. Bork P, Holm L, Sander C. 1994. *J. Mol. Biol.* 242:309-20
174. Iwanaga S, Miyata T, Tokunaga F, Muta T. 1992. *Thrombosis Res.* 68:1-32
175. Bevilacqua M, Butcher E, Furie B, Furie B, Gallatin M, et al. 1991. *Cell* 67:233
176. Drickamer K. 1993. *Curr. Opin. Struct. Biol.* 3:393-400
177. Hoppe HJ, Reid KBM. 1994. *Protein Sci.* 3:1143-58
178. Klambt C, Muller S, Lutzelschwab R, Rossa R, Totzke F, Schmidt O. 1989. *Dev. Biol.* 133:425-36
179. Exposito JY, van der Rest M, Garrone R. 1993. *J. Mol. Evol.* 37:254-59
180. Chu ML, Zhang RZ, Pan TC, Stokes D, Conway D, et al. 1990. *EMBO J.* 9:385-93
181. Rehn M, Pihlajaniemi T. 1994. *Proc. Natl. Acad. Sci. USA* 91:4234-38
182. Oh SP, Kamagata Y, Muragaki Y, Timmons S, Ooshima A, Olsen BR. 1994. *Proc. Natl. Acad. Sci. USA* 91:4229-33
183. Myers JC, Kivirikko S, Gordon MK, Dion AS, Pihlajaniemi T. 1992. *Proc. Natl. Acad. Sci. USA* 89:10144-48
184. Koller E, Winterhalter KH, Trub B. 1989. *EMBO J.* 8:1073-77
185. Krieger M, Acton S, Ashkenas J, Pearson A, Penman M, Resnick D. 1993. *J. Biol. Chem.* 268:4569-72
186. Resnick D, Pearson A, Krieger M. 1994. *Trends Biochem. Sci.* 19:5-8
187. Engel J. 1992. *Biochemistry* 31:10643-51
188. Hunter I, Schulthess T, Engel J. 1992. *J. Biol. Chem.* 267:6006-11
189. Doolittle RF. 1992. *Protein Sci.* 1:1563-77
190. Norenberg U, Wille H, Wolff JM, Frank R, Rathjen FG. 1992. *Neuron* 8:849-63
191. Yamamoto T, Gotoh M, Sasaki H, Terada M, Kitajima M, Hirohashi S. 1993. *Biochem. Biophys. Res. Commun.* 193:681-87

192. Ichijo H, Hellman U, Wernstedt C, Gonez LJ, Claesson-Welsh L, et al. 1993. *J. Biol. Chem.* 268:14505-13
193. Koschinsky ML, Beisiegel U, Henne-Bruns D, Eaton DL, Lawn RM. 1990. *Biochemistry* 29:640-44
194. Ileo K, Takahashi K, Gojobori T. 1991. *FEBS Lett.* 287:146-48
195. Pesole G, Gerardi A, di Jeso F, Saccone C. 1994. *Genetics* 136:255-60
196. Jenne D. 1991. *Biochem. Biophys. Res. Commun.* 176:1000-6
197. Kreis M, Forde BG, Rahman S, Mifflin BJ, Shewry PR. 1985. *J. Mol. Biol.* 183:499-502
198. Koonin EV, Bork P, Sander C. 1994. *EMBO J.* 13:493-503
199. Sikorski RS, Boguslawski MS, Goebel M, Hieter P. 1990. *Cell* 60:307-17
200. Doolittle RF, Woodbury NW, Jue RA. 1982. *Biosci. Rep.* 2:405-12
201. Regnier P, Grunberg-Manago M, Portier C. 1987. *J. Biol. Chem.* 262:63-68
202. Doolittle RF, Johnson MS, Husain I, Van Houten B, Thomas DC, Sancar A. 1986. *Nature* 323:451-53
203. Gorbalenya AE, Koonin EV. 1990. *J. Mol. Biol.* 213:583-91
204. Meinke A, Gilkes NR, Kilburn DG, Miller RC, Warren RAJ. 1991. *J. Bacteriol.* 173:7126-35
205. Fujii T, Miyashita K. 1993. *J. Gen. Microbiol.* 139:677-86
206. Gilkes NR, Henrissat B, Kilburn DG, Miller RC, Warren RAJ. 1991. *Microbiol. Rev.* 55:303-15
207. Ramalingam R, Blume JE, Ennis HL. 1992. *J. Bacteriol.* 174:7834-38
208. Hansen CK. 1992. *FEBS Lett.* 305:91-96
209. Little E, Bork P, Doolittle RF. 1994. *J. Mol. Evol.* 39:631-43
210. Lupas A, Engelhardt H, Peters J, Santarius U, Volker S, Baumeister W. 1994. *J. Bacteriol.* 176:1124-233
211. Engel AM, Cejka Z, Lupas A, Lottspeich F, Baumeister W. 1992. *EMBO J.* 11:4369-78
212. Foster SJ. 1993. *Mol. Microbiol.* 8:299-310
213. Guss B, Eliasson M, Olsson A, Uhlen M, Frej AK, et al. 1986. *EMBO J.* 5:1567-75
214. Kastern W, Sjobring U, Bjorck L. 1992. *J. Biol. Chem.* 267:12820-25
215. deChateau M, Bjorck L. 1994. *J. Biol. Chem.* 269:12147-51
216. Ouzounis C, Bork P, Sander C. 1994. *Trends Biochem. Sci.* 19:199-200
217. Saier MH, Reizer J. 1990. *Res. Microbiol.* 141:1033-38
218. Doolittle RF. 1994. *Trends Biochem. Sci.* 19:15-18
219. Goundis D, Reid KBM. 1988. *Nature* 335:82-85
220. Robson KJH, Hall JRS, Jennings MW, Harris TJR, Marsh K, et al. 1988. *Nature* 335:79-82
221. Pereira MEA, Mejia JS, Ortega-Barria E, Matzilevich D, Prioli RP. 1991. *J. Exp. Med.* 179-91
222. Blomquist MC, Hunt LT, Barker WC. 1984. *Proc. Natl. Acad. Sci. USA* 81:7363-67
223. Kotwal GJ, Moss B. 1988. *Nature* 335:176-78
224. Beckmann G, Bork P. 1993. *Trends Biochem. Sci.* 18:40-41
225. Jank B, Habermann B, Schweyen RJ. 1993. *Trends Biochem. Sci.* 18:427-28
226. Reizer J, Reizer A, Saier MH. 1993. *Trends Biochem. Sci.* 18:247-48
227. Kwong PD, McDonald NQ. 1993. *Trends Biochem. Sci.* 18:425-26
228. Creighton TE, Kimmink J. 1993. *Trends Biochem. Sci.* 18:424-25
229. McDonald NQ, Kwong PD. 1993. *Trends Biochem. Sci.* 18:208-9
230. Petrou S, Ordway RW, Singer JJ, Walsh JV. 1993. *Trends Biochem. Sci.* 18:41-42
231. Grandori R, Carey J. 1994. *Trends Biochem. Sci.* 19:72
232. Koonin EV, Bork P. 1994. *Trends Biochem. Sci.* 19:234-35
233. Chothia C. 1992. *Nature* 357:543-44
234. Blundell TL, Johnson MJ. 1993. *Protein Sci.* 2:877-83
235. Murzin AG, Chothia C. 1992. *Curr. Opin. Struct. Biol.* 2:895-903
236. Farber GK. 1993. *Curr. Opin. Struct. Biol.* 3:409-12
237. Doolittle RF, Blundell TL. 1993. *Curr. Opin. Struct. Biol.* 3:377-78