PROTEIN FAMILIES AND THEIR EVOLUTION—A STRUCTURAL PERSPECTIVE

Christine A. Orengo¹ and Janet M. Thornton²

¹Department of Biochemistry and Molecular Biology, University College, London WC1E 6BT, United Kingdom; email: orengo@biochemistry.ucl.ac.uk
²European Bioinformatics Institute, Hinxton Campus, Cambridge CB10 1SD, United Kingdom; email: thornton@ebi.ac.uk

Key Words protein classifications, comparative genomics, bioinformatics

Abstract We can now assign about two thirds of the sequences from completed genomes to as few as 1400 domain families for which structures are known and thus more ancient evolutionary relationships established. About 200 of these domain families are common to all kingdoms of life and account for nearly 50% of domain structure annotations in the genomes. Some of these domain families have been very extensively duplicated within a genome and combined with different domain partners giving rise to different multidomain proteins. The ways in which these domain combinations evolve tend to be specific to the organism so that less than 15% of the protein families found within a genome appear to be common to all kingdoms of life. Recent analyses of completed genomes, exploiting the structural data, have revealed the extent to which duplication of these domains and modifications of their functions can expand the functional repertoire of the organism, contributing to increasing complexity.

CONTENTS

INTRODUCTION ......................................................... 868
CLASSIFYING AND CHARACTERIZING PROTEIN FAMILIES .............. 871
Families and Superfamilies .............................................. 871
Fold Groups ............................................................. 872
Domain Structure Architectures and Classes ............................. 872
Orthologous and Paralogous Relatives .................................. 873
HOW DO WE FIND PROTEIN RELATIVES? ............................. 873
Finding Close Relatives .................................................. 873
Finding Distant Relatives in the Twilight Zone ......................... 875
Finding Very Distant Relatives in the Midnight Zone ................. 875
WHAT DO THE CURRENT CLASSIFICATIONS REVEAL ABOUT
THE NUMBER AND NATURE OF DOMAIN FAMILIES? .................. 878
How Many Domain Families Can Be Identified in Sequence Databases? 878
How Many Domain Families Are Currently Identified Using Structural Data? 880
Are the Domain Families and Folds Equally Populated? ............ 880
Are Folds Distinct or Is There a Structural Continuum? ............... 881

0066-4154/05/0707-0867$20.00
INTRODUCTION

The unraveling of the genetic code by Watson and Crick, over 50 years ago, started a new era in evolutionary biology. Building on these insights came the revolutionary technologies for sequencing proteins, developed by Sanger in the early 1950s. These were quantum leaps in biology, and the resulting expansions in the datasets of known protein sequences by the international genome projects, together with significant advances in the computational methods for detecting similarities between evolutionarily related genes, are now promising to yield profound insights into the evolution of proteins, their functions, and the biological processes in which they participate.

The mechanisms by which genomic DNA can change during evolution are now being elucidated, thanks to the explosion of data from these sequencing projects and the growing diversity of genomes from all kingdoms of life. Many proteins in these organisms comprise more than one domain (for example, see Figure 10, below). Although the importance of domain duplication in evolution has long been recognized, analyses of completed genomes have confirmed the extent to which this duplication is clearly occurring (1). In prokaryotes, at least 70% of the domains have been duplicated, whereas in eukaryotes this figure appears to be as high as 90% (2).

Computational analyses of data from both prokaryotic and eukaryotic genomes, reviewed below, have confirmed the importance of the protein domain as a fundamental unit in evolution and revealed the astonishing diversity of proteins that
can be assembled by duplicating domains and then combining them in different ways (1, 2). Some domain families appear to be intrinsically more versatile, recurring much more frequently in the genomes with many different domain partners. Others, by contrast, currently appear unique to the organism or kingdom, often occurring as single domains or with only 1 or 2 different domain partners (1–4).

Not long after discovery of DNA and the development of protein sequencing technologies, methods for determining the 3-D structures became established in the late 1970s. These methods allowed biologists to inspect and probe the interactions between amino acid residues in a protein that determine the fold and the manner in which proteins interact with other proteins and substrates in their environment. As the number of known structures solved by X-ray crystallography and NMR techniques increased, it became clear that protein structure is much more highly conserved throughout evolution than the protein’s sequence (5) (Figures 1 and 2). In contrast to the protein sequence, where in some families relatives have been detected sharing fewer than 5% identical residues, in many protein families at least 50% of the structure, mainly in the core of the protein, is highly conserved (6, 7) and can be used as a fingerprint to detect very distant relatives (8).

Thus, although many excellent sequence-based resources have been developed over the past 20 years that classify and characterize protein families (9, 10), the

![Figure 1](https://example.com/figure1.png)

**Figure 1** Correlation between structure similarity (measured by the SSAP structure comparison algorithm, 0–100) and sequence similarity (measured by sequence identity) for pairs of homologous domain structures in the CATH domain database. Homologous proteins possessing the same function are labeled as circles. Squares indicate homologous relatives with different functions.
Figure 2  Schematic representation of the progression from close homologues, through more remote (twilight zone) and very remote (midnight zone) homologues and finally analogous structural relatives.
more recent structure-based resources often allow us to recognize more ancestral relationships. Sometimes this structural data provides a clearer picture of how evolutionary processes have exploited the domain family repertoire to build new domain combinations. These mechanisms appear to play a major role in increasing the complexity of organisms, thereby giving rise to the resulting diversity of phenotypes observed in nature (11–13).

In this review, we consider the challenges faced in recognizing and classifying evolutionary relatives and discuss how these problems have been partly addressed by significant improvements in computational approaches for homologue recognition. We briefly review some of the major sequence-based family classifications before considering the extra sensitivity that can be achieved in homologue recognition using structural data, when available. We then describe the major resources that classify structural families. Because these resources can capture information on more ancient evolutionary relationships, we focus primarily on what we can learn about the evolution of protein families and their functions by exploiting these structural classifications.

In particular, we review several interesting discoveries about protein family distributions across all kingdoms of life and speculate on what these analyses reveal about the most common, and therefore probably the most ancestral, protein families. We also consider how expansions in these ancient families may have contributed to the complexity and diversity of life. For many of the domain families that are highly recurrent in the genomes, structural data are starting to provide profound insights into the mechanisms by which domain duplication followed by divergence and/or domain fusion events have modulated the functions of the proteins.

CLASSIFYING AND CHARACTERIZING PROTEIN FAMILIES

A large proportion of genes, up to 90% in eukaryotes, compose multidomain proteins (1). Thus, in some sense the domain can be viewed as a primary unit of evolution. Therefore, in our description of protein family classifications, we concentrate primarily on domain classifications before considering how domains are duplicated and combined in various ways to give different protein families.

Before describing protein classification strategies and analyses, we first define the terms most frequently used in connection with protein families and with groups of related proteins deriving from both divergent and convergent evolution. The concepts can perhaps be more clearly understood by following the different levels in the hierarchy illustrated in Figure 2.

Families and Superfamilies

During the course of evolution, proteins derived from a common ancestral protein can change their sequences and diverge by mutations or substitutions of the residues and also by insertions and deletions of residues (indels), giving rise to families of
homologous proteins. Not all positions are equally susceptible to mutation as some positions may be very important for function, stability, or folding and may thus be more constrained in the residue types allowed.

Many protein family resources present a hierarchical classification whereby very close relatives, for example with high sequence similarity (e.g., >40% sequence identity), are grouped together into families. These close relatives frequently share common functional properties. More remote homologues that have lower sequence similarity (<30%) are grouped together into broader evolutionary families or superfamilies, a term first coined by Margaret Dayhoff (13a) based on her recognition of the extent to which proteins could diverge within a family.

Figure 2 shows an example of close and remote relatives within the P-Loop Hydrolase family. It is difficult to recognize very divergent relatives by comparing their sequences alone, and in the example shown in Figure 1, the remote homologues could only be detected by comparing their structures. Further evidence of their evolutionary relationship came from an understanding of the common mechanisms by which they performed their different functions.

Confusion can sometimes arise in deciding whether a group of homologues represents a close family or a broader superfamily. Therefore, to simplify our discussions, here we use the term family in the broadest sense, i.e., containing all the relatives, orthologues, and paralogues.

Fold Groups

Sometimes structural similarity is apparent between two domains that share no other common characteristic. In these cases, where the relatives possess no significant sequence similarity and no features indicative of common functional properties, it is very difficult to determine the evolutionary relationship. They are possibly two extremely distant relatives from the same evolutionary family that have diverged far in both sequence and function. Another possibility is that they have come from different ancestral proteins or domains but have both converged on the same structural arrangement or fold.

Protein structures are primarily composed of α-helical and β-strand secondary structures, and there are clearly restraints on the ways in which these can be packed together to achieve optimal packing of hydrophobic residues in the cores of the proteins (14). Domains or proteins sharing structural similarity but no common sequence or functional characteristic are described as analogues. For these proteins, there is some ambiguity as to whether they are related by convergent evolution to a favorable structure or divergent evolution from a common ancestor. Figure 2 shows an example of homologous and analogous structures adopting a Rossmann-like fold.

Domain Structure Architectures and Classes

As well as recognizing distinct folds, most structure classifications also describe the architecture of the protein structure (15, 16). This is really the overall shape
of the protein. Whereas topology or fold group describes the relative orientations of the secondary structures in 3D and the order in which they are connected, the architecture is a higher level in the classification and groups together structures with similar secondary structural arrangements regardless of connectivity (see Figure 3).

At the highest level in a structural classification, proteins are grouped if they belong to the same class, i.e., if they have similar secondary structure compositions and packing. There are three major classes, mainly $\alpha$, mainly $\beta$, and $\alpha-\beta$, although this latter class can be subdivided into alternating $\alpha/\beta$ in which $\alpha$ and $\beta$ secondary structures alternate along the protein chain and $\alpha+\beta$ in which the $\alpha$ and $\beta$ regions are largely segregated.

Orthologous and Paralogous Relatives

Domain and gene duplication has clearly been an important and frequent feature of evolution (1). Following duplication, the original domain tends to retain its function, especially where this is important to the integrity of the organism. However, the duplicated domain has no such functional constraints and may more likely be retained if it brings an additional and beneficial new function to the organism. Duplicated genes or domains are described as paralogues, and in these relatives the sequences and functions can diverge considerably from the original domain (see Figure 2).

By contrast, the term orthologues describes related domains and genes that have diverged from the same common ancestral gene by a process of speciation. Orthologous relatives usually possess very similar functions (for more detailed explanations, see Reference 17).

HOW DO WE FIND PROTEIN RELATIVES?

Finding Close Relatives

As the sequence databases expanded during the early 1970s, various means were devised for aligning and comparing the sequences. The challenge is to recognize equivalent residues despite the noise of the sometimes very extensive insertions that occur between distant relatives. Needleman & Wunsch found an elegant solution to this challenge based on dynamic programming algorithms (18), a branch of optimization theory that has been exploited largely for solving the traveling salesman and other related problems. Various implementations of this approach were adopted to find both local (19) and global (18, 20) similarities between proteins, and dynamic programming is still used extensively by most protein family classifications to detect equivalent positions between closely related proteins (7, 18, 20) and measure their sequence identities.

Dynamic programming methods are computationally demanding. With the exponential increases in the sequence databases, other methods were designed that
Figure 3  The four major hierarchical levels in the CATH structural classification: class, architecture, topology or fold level, and homologous superfamily. Three of the most highly populated architectures in the classification are illustrated.
approximated these approaches by searching for matching fragments between proteins, thereby ignoring the insertions [FASTA (21), BLAST (22)]. This simplification allowed the databases to be searched 2 or 3 orders of magnitude faster, which meant that millions of comparisons could be performed, providing a large dataset of random alignments and similarities to be calculated. This dataset gives a robust statistical framework for assessing the significance of any putative match between two protein sequences (22).

Finding Distant Relatives in the Twilight Zone

However, as the number of protein 3-D structures deposited in the Protein Databank (PDB) (23) grew, it became clear that some homologues could diverge to a point of insignificant sequence similarity (see examples in Figure 2). Pair-wise sequence comparison methods have difficulty in recognizing these relatives, even using substitution matrices that model allowed residue mutations (20). For relatives in this twilight zone of sequence similarity coined by Feng & Doolittle (23a), homologue recognition was improved by building sequence profiles using a set of known homologues from a protein family to determine the constraints on evolutionary change (20, 24).

Family specific residue patterns or profiles are built by automatically analyzing a multiple alignment of the sequence relatives, thereby characterizing residue preferences at each position in the protein [e.g., PsiBLAST (25)]. Some protein family resources also use multiple alignments to derive regular expressions of highly conserved sequence motifs that can be used as a fingerprint of a particular family to recognize distant homologues (26).

More recently, a new family of algorithms, Hidden Markov Models (e.g., HMMer, SAM-T; 27–29), have been shown to be highly sensitive in recognizing very remote homologues, some of which can be considered to occupy the midnight zone of sequence similarity, recently coined by Rost (see Figure 2). These algorithms exploit more robust solutions for modeling observed residue insertions and deletions across the family, and some approaches have been shown to be up to four times more powerful than BLAST (28, 29). Table 1 lists some of the most comprehensive and well-established protein family databases, in which relatives have been classified primarily using sequence-based methods.

Finding Very Distant Relatives in the Midnight Zone

To trace further back in evolution into the midnight zone of sequence similarity, it is often necessary to exploit structural data in some way. A new breed of algorithms, described as threading, were developed in the early 1990s, pioneered by Jones et al. (30), Bowie et al. (31), and other groups (32). Threading attempts to thread a protein sequence through representatives of known structural families to recognize the most energetically favorable fit (for more details, see review in 32). As these approaches can be very slow, they are not generally used for assigning large sets of sequences (e.g., from genomes) to protein families, although some
TABLE 1  Protein sequence family resources

<table>
<thead>
<tr>
<th>Resource</th>
<th>Group</th>
<th>Source(s)</th>
<th>Number of families</th>
<th>Method</th>
<th>URL</th>
</tr>
</thead>
<tbody>
<tr>
<td>PRINTS</td>
<td>Zygouri</td>
<td>SWISSPROT, TrEMBL</td>
<td>1,800 entries, 10,931 motifs</td>
<td>Iterative motif searches</td>
<td><a href="http://bioinf.man.ac.uk/dbbrowser/PRINTS/">http://bioinf.man.ac.uk/dbbrowser/PRINTS/</a></td>
</tr>
<tr>
<td>Pfam</td>
<td>Eddy</td>
<td>SWISSPROT, TrEMBL</td>
<td>7,459 families</td>
<td>HMM</td>
<td><a href="http://www.sanger.ac.uk/Software/Pfam">http://www.sanger.ac.uk/Software/Pfam</a></td>
</tr>
<tr>
<td>SMART</td>
<td>Bork</td>
<td>Selected proteins</td>
<td>667 domains</td>
<td>HMM</td>
<td><a href="http://smart.embl-heidelberg.de">http://smart.embl-heidelberg.de</a></td>
</tr>
<tr>
<td>InterPro</td>
<td>Zbobnov</td>
<td>UniProt, PROSITE, PRINTS, Pfam, ProDom, SMART, TIGRFAMs, PIR, SuperFamily, SUPERFAMILY</td>
<td>11,007 entries (including 2,573 domains, 8,166 families)</td>
<td>Multiple methods (HMM, PSI-BLAST regular expression)</td>
<td><a href="http://www.ebi.ac.uk/interpro">http://www.ebi.ac.uk/interpro</a></td>
</tr>
<tr>
<td>TIGRFAMs</td>
<td>White</td>
<td>SWISSPROT, TrEMBL</td>
<td>1,976 families</td>
<td>HMM</td>
<td><a href="http://www.tigr.org/TIGRFAMs/index.shtml">http://www.tigr.org/TIGRFAMs/index.shtml</a></td>
</tr>
<tr>
<td>ADDA</td>
<td>Holm</td>
<td>SWISSPROT, TrEMBL, PIR, PDB, WORMPEP, ENSEMBL</td>
<td>34,000 families (plus 60,000 singleton)</td>
<td></td>
<td><a href="http://ekhidna.biocenter.helsinki.fi:8080/examples/servlets/adda/index.html">http://ekhidna.biocenter.helsinki.fi:8080/examples/servlets/adda/index.html</a></td>
</tr>
<tr>
<td>CHOP</td>
<td>Rost</td>
<td>62 complete genomes</td>
<td>63,300 clusters (plus 118,108 singleton clusters)</td>
<td>PSI-BLAST</td>
<td><a href="http://cubic.bioc.columbia.edu/services/CHOP">http://cubic.bioc.columbia.edu/services/CHOP</a></td>
</tr>
<tr>
<td>Database</td>
<td>Creator</td>
<td>Description</td>
<td>Search Options</td>
<td>Website</td>
<td></td>
</tr>
<tr>
<td>------------</td>
<td>-------------</td>
<td>-------------------------------------------------------------</td>
<td>--------------------------------------------------------------------------------</td>
<td>----------------------------------------------</td>
<td></td>
</tr>
<tr>
<td>TRIBES</td>
<td>Ouzounis</td>
<td>83 complete genomes</td>
<td>60,934 or 82,692 depending on granularity</td>
<td><a href="http://maine.ebi.ac.uk:8000/services/tribes">http://maine.ebi.ac.uk:8000/services/tribes</a></td>
<td></td>
</tr>
<tr>
<td>ProtoNet</td>
<td>Linial</td>
<td>SWISSPROT, TrEMBL</td>
<td>User defined</td>
<td><a href="http://www.protonet.huji.ac.il">http://www.protonet.huji.ac.il</a></td>
<td></td>
</tr>
<tr>
<td>SYSTERS</td>
<td>Vingron</td>
<td>SWISSPROT, TrEMBL, ENSEMBL (complete genomes), the Arabidopsis Information Resource, SGD, and GeneDB</td>
<td>158,153 disjoint clusters</td>
<td><a href="http://systers.molgen.mpg.de">http://systers.molgen.mpg.de</a></td>
<td></td>
</tr>
<tr>
<td>iProClass</td>
<td>Wu</td>
<td>PIR, SWISSPROT, TrEMBL, PFam, BLOCKS, PRINTS, ProSite, PDB, COG</td>
<td>36,000 PIR superfamilies, 100,000 families</td>
<td><a href="http://pir.georgetown.edu/iproclass">http://pir.georgetown.edu/iproclass</a></td>
<td></td>
</tr>
<tr>
<td>SWISSPROT</td>
<td>Schneider</td>
<td>Primary database</td>
<td>153,871 proteins</td>
<td><a href="http://us.expasy.org/sprot">http://us.expasy.org/sprot</a></td>
<td></td>
</tr>
</tbody>
</table>
recent implementations and approximations combine speed with some increased sensitivity (33–35).

Most approaches that rely on sequence data still fail to recognize up to 30% of extremely distant homologues (4, 29, 36). For these cases, only direct structure comparison reveals the evolutionary relationship. More than 50 methods have been devised over the past 15 years (7, 37). The earliest methods, which exploited rigid-body superposition of structures, often have difficulty handling extensive indels between distant homologues. It was the adaptation of dynamic programming algorithms to structure alignment in the late 1980s (7, 38, 39, 41, 45) that allowed recognition of very remote homologues. Other optimization algorithms have also been exploited [e.g., Monte Carlo optimization (40), simulated annealing (39)].

However, as with threading, most methods that compare structural properties of residues are slow and not suited to large-scale comparisons. Also exploited are rapid comparison protocols comparing secondary structures between proteins, of which there are an order of magnitude fewer than residues (42–44). These protocols have the advantage of enabling a large number of comparisons to be made between unrelated proteins to assess the statistical significance of any similarity.

As well as enabling the evolutionary fingerprints between very distant homologues to be detected, structural data allow more facile recognition of the location of domain boundaries in multidomain proteins. Various algorithms have been developed to optimize domain boundary recognition (46–51).

Table 2 summarizes the most comprehensive and widely accessed domain structure classifications and the populations of these resources, as of August 2004. Most of the databases exploit automated structure comparison approaches. However, structural similarity alone is not sufficient to prove evolutionary ancestry, and in the major structural classifications (e.g., SCOP, CATH), additional evidence (e.g., common sequence patterns, functional similarity) is sought, sometimes by extensive manual validation. By contrast, the structural neighborhood resources set up by various bioinformatics institutes [DDD (52), CE (45), SSM (44), CDD (53)] automatically detect close or distant structural similarities between proteins without formally classifying relatives into families.

WHAT DO THE CURRENT CLASSIFICATIONS REVEAL ABOUT THE NUMBER AND NATURE OF DOMAIN FAMILIES?

How Many Domain Families Can Be Identified in Sequence Databases?

How many domain families can be identified in the current set of known protein sequences and structures? Obviously, the sequence data are much more comprehensive, and a number of approaches have been developed that perform extensive sequence comparisons (e.g., using BLAST or PsiBLAST) and then exploit the...
### TABLE 2  Protein structure family resources

<table>
<thead>
<tr>
<th>Database</th>
<th>Location and author</th>
<th>Coverage (July 2004)</th>
<th>Structure comparison method</th>
<th>Type</th>
<th>URL</th>
</tr>
</thead>
<tbody>
<tr>
<td>CAMPASS</td>
<td>Cambridge University, UK Sowdhamini</td>
<td>7,580 domains in 1,409 superfamilies</td>
<td>COMPARER (39), SEA (103)</td>
<td>Structure-based sequence alignments of SCOP superfamilies</td>
<td><a href="http://www-cryst.bioc.cam.ac.uk/~campass/">http://www-cryst.bioc.cam.ac.uk/~campass/</a></td>
</tr>
<tr>
<td>CATH Gene3D</td>
<td>UCL, London, UK Orengo</td>
<td>58,000 domains in 1,459 superfamilies</td>
<td>SSAP (38), GRATH (43)</td>
<td>Automatic structural and sequence comparison methods are combined with manual validation of superfamily alignments and domain boundaries</td>
<td><a href="http://www.biochem.ucl.ac.uk/bsm/cath/">http://www.biochem.ucl.ac.uk/bsm/cath/</a></td>
</tr>
<tr>
<td>CE</td>
<td>SDSC, La Jolla, CA, USA Bourne</td>
<td>All chains in PDB</td>
<td>CE (45)</td>
<td>Fully automatic Nearest neighbors</td>
<td><a href="http://cl.sdsc.edu/ce.html">http://cl.sdsc.edu/ce.html</a></td>
</tr>
<tr>
<td>DHS</td>
<td>UCL, London, UK</td>
<td>1,459 superfamilies in CATH</td>
<td>SSAP (38), CORA (8)</td>
<td>Fully automatic multiple structure alignments of close relatives in CATH superfamilies</td>
<td><a href="http://www.biochem.ucl.ac.uk/bsm/dhs/">http://www.biochem.ucl.ac.uk/bsm/dhs/</a></td>
</tr>
</tbody>
</table>
notion of domain recurrence to group complete or partial sequences into domain families. This is a difficult task as reflected in the significant variations in numbers of families identified \[\sim 20,000 - \sim 60,000\] (4, 54–57).

A better idea of the number of families can be obtained by using more powerful approaches, such as HMMs, to improve the detection of remote homologues. Pfam (58), the most comprehensive domain resource based on sequence data, currently contains about 7600 domain families that account for a significant proportion (up to 80%) of sequences in completed genomes, depending on the organism (57).

A clearer grasp of the total number of domain families will emerge in the near future, helped enormously by the InterPro resource (10), which is integrating all the domain families from established databases [e.g., Pfam (58), PRINTS (26), SMART (59), Tigrfam (60), SCOP (61), CATH (62)] and mapping them onto the genome sequences.

### How Many Domain Families Are Currently Identified Using Structural Data?

Structural data allow even more distant relationships to be found, further collapsing the number of families. For example, Pfam families for which structures are known can be merged if their structures are classified in the same CATH family. This merger gives a nearly twofold collapse in the number of families characterized in this way (57).

We consider the nature and distribution of the structural families in more detail because they are perhaps the best characterized of domain families; the conservation of the structural fingerprint ensures even very distant relatives are recognized. The accuracy in recognizing domain boundaries from structural data also means that these families are likely to constitute the most accurate domain dictionary. Currently, between 1200 and 1400 domain families are classified in the SCOP (16) and CATH (15) databases, and nearly 70% of domain sequences in the genomes can be assigned to these families using the most powerful sequence comparison methods (4).

### Are the Domain Families and Folds Equally Populated?

Among the 1400 structural families currently classified in CATH, we find fewer than 850 different folds (see Figure 3) (62). A very small percentage of these fold groups (<10 folds, <0.1% of the total) are very large, accounting for nearly 40% of all sequence families in the PDB (62). All of these very common folds are adopted by several different superfamilies, whereas most of the remaining folds in CATH (~85%) correspond to a single evolutionary family. When this feature was first observed in the mid-1990s, these popular folds were referred to as superfolds (63, 64). It can be seen from the CATH wheel shown in Figure 4 that some of these superfolds are adopted by more than 10 different homologous superfamilies, many of which are also highly populated with diverse sequences.

This uneven distribution of folds in the PDB is also seen at higher levels in the classification. Although there about 30 architectural types defined in the CATH
classification (62, 65), a handful of these [$\alpha\beta$-sandwiches (two- and three-layer), $\alpha\beta$-barrel, $\beta$-barrel, $\alpha$-updown] contain 30% of the fold groups and half the superfamilies (65). These are the most regular architectures in the classification comprising layers of $\alpha$-helices and $\beta$-strands. The three-layer $\alpha\beta$-sandwich shown in Figure 2 is adopted by nearly 80 different folds.

The most highly populated fold groups or superfolds adopt these very regular architectures (e.g., TIM barrel fold, $\alpha\beta$-barrel, Rossmann fold; three-layer, $\alpha\beta$-sandwich; $\alpha\beta$-plait, two-layer $\alpha\beta$-sandwich). Such 3-D arrangements might be expected to produce more favorable residue packing and also enable the structures to maintain this packing following residue substitutions, as secondary structure layers could slide more easily relative to each other, thereby accommodating changes in the sizes or shapes of residues.

This characteristic may contribute to the preponderance of the superfolds, whether or not these fold groups reflect convergent or divergent evolution. The regularity and optimal packing of the fold would ensure that more protein structures converged on these arrangements during evolution. Conversely, the ability to sustain residue substitutions would allow the sequences to change considerably by divergent evolution. Recent statistical analyses exploiting energetic considerations have shown that the superfolds are able to support a much broader repertoire of sequences than other folds (66). In support of divergent evolution, recent analyses of 18 diverse superfamilies adopting the TIM barrel fold found plausible evidence for homologous relationships between several of these different families (67).

Are Folds Distinct or Is There a Structural Continuum?

Even more speculative are suggestions that perhaps the earliest evolutionary unit corresponded to a much smaller structural motif than a domain (68, 69) (e.g., a supersecondary structural motif such as an $\alpha\beta$-motif, $\beta$-hairpin, $\alpha$-hairpin). This idea lends support to the notion of a structural continuum proposed by several researchers (15, 70) to explain the observation that some regions of fold space are very densely populated so that distinguishing between different folds becomes difficult and subjective (71).

For example, many folds possess $\alpha\beta$-sandwich and $\beta$-sandwich architectures (nearly one third of all fold groups, see Figure 3), and folds adopting these architectures often share large common structural motifs due to recurrence of simple motifs (e.g., $\beta\beta$ motifs, $\alpha\beta$ motifs, split $\beta\alpha\beta$ motifs). In some cases, structural similarity between overlapping fold groups is as high (up to 50% overlap) as the similarity displayed across a very diverse evolutionary superfamily (Figure 5) (6, 43).

How Many Domain Families and Folds Could There Be in Nature and How Evenly Are They Distributed?

The completion of the human genome at the turn of the century, followed over the past five years by 200 other genomes, including 15 eukaryotes (e.g., human, mouse, fly, worm), allows much better insights into the distribution and recurrence
of domain families in nature. Can we use these genome data to predict the total number of domain families or even folds in nature?

Nearly a decade ago, when fewer than 300 folds were characterized, there was speculation on the probability of there being fewer than 1000 folds in nature (63, 72). By considering the extent to which sequence families merge once structural data are acquired, it is possible to estimate the number of unique structures in nature, assuming that most sequence families sample randomly from fold space.
However, the numbers calculated vary considerably, from a few thousand to tens of thousands, depending on the models applied (73–76).

If we map relatives from the domain family resources, CATH and Pfam, onto the genome sequences, using powerful sequence comparison methods (57), the majority of domain sequences from the genomes (>70%) can be assigned to fewer than 2500 of the largest families from these two resources (Figure 6). Figure 7 summarizes the organization of sequence data from the completed genomes into domain families using these sequence-based and structure-based classifications.

The other trend to emerge very clearly from the genome data is the extreme bias in the distributions of domain families, whereby a very small proportion of families recurs very extensively within genomes although most families occur only once or a few times within a specific organism or subkingdom (1, 104, 105). These power-law-like trends are illustrated in Figure 8 (a–c), which displays the recurrence of domain relatives from CATH and Pfam in the genome sequences. For both types of domains a small percentage of families (<10%) are recurring...
Figure 6  Proportion of domain sequences in completed genomes that can be assigned to the domain families in the CATH and Pfam domain databases. The x-axis is labeled by the number of families in CATH and Pfam in order of the size of the family, with the largest families labeled first.

tens or hundreds of times within a single genome. The structure-based families from CATH appear to recur more extensively that those characterized in Pfam, but this simply reflects our ability to track more distant relatives using structure.

Partial genome sequences not assigned to CATH and Pfam are assigned to about 45,000 very small families (<5 relatives in each), which are specific to the organism or subkingdom (57). These are mostly small proteins (<200 residues) and thus probably constitute single domain structures. These unassigned families may therefore represent a large, indeterminate tail of structures specific to individual organisms.

Which Are the Most Ancient Domains Common to All Kingdoms of Life and What Are Their Roles?

The current complement of genomes, from all kingdoms of life, allows us to search for common families and explore the emergence of new domains during evolution. That this is timely is demonstrated by recent comparisons of domain family repertoires between organisms used to derive phylogenetic trees. Encouragingly, these trees have shown remarkable resemblance to the RNA-derived trees (77).

Figure 5  Illustration of structural overlaps between proteins (a) from the same homologous superfamily, P-loop hydrolases, (b) from the same homologous superfamily, the galactin-type carbohydrate recognition domain superfamily, and (c) between proteins adopting different folds in the mainly β-sandwich architecture. The PDB code and CATH identification number are shown beneath each structure.
Using sequence data alone, we find that the most common domains are primarily involved in protein biosynthesis (78). This is not surprising as these are critical functions for all organisms. However, if structural data are exploited, a more complex picture emerges (12, 13, 57, 76). We can detect about 200 domain structure families common to all kingdoms of life, and among this set we also find families involved in metabolism and regulation (57).

Some of these families are very extensively duplicated during evolution, accounting for a significant proportion of domain structure sequences within a genome (up to 50% in bacteria, see also below). However, subsequent divergence means that little sequence signal remains as fingerprints of their evolutionary relationship. Only structural data allow us to recognize this common
ancestry. These data now enable us to piece together some of the earliest pathways and processes present in the last common universal ancestor (LUCA) of all species.

What domains and folds have evolved more recently during evolution? Again, by considering the structural data, we can see different trends emerging in the different lineages. Several new mainly $\beta$-folds appear to have arisen in the eukaryotes. More mainly $\alpha$-folds have emerged in the bacterial lineages, and there have been fewer new $\alpha\beta$-folds in any of the kingdoms (79).

We should treat these data with some caution, however. The apparent emergence of new folds may be an artifact caused by our inability to detect extremely diverged structures. The $\alpha/\beta$ architectures mainly comprise regular barrel and sandwich-like shapes, e.g., TIM barrels and Rossmann-like folds, of which some are very highly

**Figure 8** Power-law-like behavior of families from (a) CATH domain database, (b) Pfam domain database, (c) uncharacterized domain families (NewFam), and (d) protein families from the Gene3D database.
populated (see Figure 4). Embellishments of these structures are mostly associated with extensions of the $\beta$-sheet, so that comparisons between very distant homologues can still identify large conserved motifs. For some irregular mainly $\beta$ and mainly $\alpha$ structures, divergence would result in more profound structural changes, making it harder to recognize the core similarity and evolutionary ancestry of these proteins.

HOW MANY PROTEIN FAMILIES CAN WE IDENTIFY IN THE GENOMES?

As mentioned, many proteins in the genomes are apparently multidomain [at least 90% in eukaryotes (1)]. So far we have only considered the number and distribution of domain families identified in the genomes. How do these numbers contrast with the number of protein families identified? In this case, families will comprise protein sequences sharing the same multidomain composition (see Figure 7).

If the protein sequences from all the completed genomes (nearly one million sequences) are compared and then clustered in a manner that tends to group proteins sharing common domain compositions, around 50,000 different protein families can be recognized (57). We have classified these families in a resource called Gene3D (57), but many other excellent resources of a similar type exist [e.g., TRIBES (80), SYSTERS (81), ProtoNet (82), COGS (83), KOGS (84)].

What Is the Distribution of the Protein Families in the Genomes?

The distribution of these protein families exhibits the same power-law-like behavior as the domain families (see Figure 8d). Furthermore, about 150,000 sequences, corresponding to between 10% and 20% of the proteins in each genome, appear to be singletons, i.e., they have no relative either within their own genome or any other genome (see Figure 9a). It is not clear whether these sequences are unique or whether they have diverged so significantly during evolution that any common ancestry with other sequences is lost.

A significant proportion are very small proteins that appear to have low secondary structure content when prediction algorithms are applied (85). This may suggest that they are associated with signaling or regulatory events in the cell and perhaps only adopt stable structures on binding to other proteins or complexes. They may therefore possess regulatory functions that are highly specific for the organism.

In analyzing the protein family data in Gene3D, we observe that less than 15% of protein families (i.e., domain combinations) in each genome are common to other genomes and kingdoms (see Figure 9a). More than two thirds of the protein families are unique to the organism and/or kingdom (see Figure 9a) (57).
Characterizing the Domain Compositions of the Protein Families

Can we determine the domain compositions of these unique proteins and search for similarities in the domain components from which they are assembled? As noted, a significant proportion of the domain sequences in the genomes can be
assigned to CATH and Pfam domain families (up to 80% of the domain sequences in some genomes). These data have been captured in our Gene3D resource (57), allowing us to explore the variety of domain compositions occurring in nature and the different ways in which the domains have been combined, giving rise to new protein families that are specific to the organism or subkingdom (see Figure 7).

Perhaps one of the most striking observations that emerges from the data is that although only 15% of protein families are common across the genomes and two thirds are specific to the organism or kingdom, by looking at the domain mapping data we can see that many protein families comprise domains that are common to all kingdoms of life (see Figure 9b). Some 50% of domain structure annotations in each organism are from fewer than 200 domain families common to all kingdoms of life.

These data also reveal that the domain family repertoire has not been exploited as fully as it could be. Although some 5000 CATH plus Pfam domain families can be mapped onto the genomes, we only observe approximately 50,000 protein families, which is much less than the $>10^7$ possible two-domain proteins or $>10^{11}$ possible combinations of three-domain proteins. Chothia and coworkers have also estimated that less than 0.5% of all possible combinations are observed (2).

The protein repertoire is smaller than possible because although many domains have been duplicated [nearly 60% in prokaryotes, at least 90% of domains in eukaryotes, and 95% in human (1, 3)], only a small proportion of these domains (<1%) have been duplicated very extensively. These highly recurrent domains are dominating the genomes, as revealed in the power-law plots described earlier (Figure 8). Many of those recurring most extensively adopt one of the ten superfolds shown in Figure 4b (e.g., the αβ-plait or Rossmann-like folds) and account for nearly 40% of the annotations.

Most of these highly recurring domains are performing important generic functions, for example providing energy or redox equivalents for reactions, or they are associated with regulation (1). Therefore, this extensive recurrence of domains is most likely due to the reuse of functionally important domain modules (3) rather than the predominance of energetically favoured folds. Not surprisingly, a bias is also observed in domain combinations whereby most domains are only observed in one or two different contexts, whereas these very highly recurrent domains have many domain partners, some more than 20 (1, 57).

These highly duplicated domains have therefore combined in different ways with many different domain partners, giving rise to the incredible variety of multidomain proteins observed in different species and kingdoms. Proteins with different domain compositions tend to have different functions (86, 88–91; see below), and so this variety in multidomain composition must be partly responsible for the diverse functions and phenotypes observed in the different organisms.

Are Some Domain Combinations More Frequent?

As well as more recurrent single domains, there are also some frequently recurring pairs of domains or supradomains (3, 92) (Figure 10). More than one third of
all proteins contain a supradomain, and they occur in all kingdoms of life and categories of function. The 200 supradomains identified to date have been found in 75,000 genome sequences (92). The manner of their recurrence suggests that they have been duplicated as one evolutionary unit, either because it is functionally beneficial to have the activities from both domains present on one protein or because the functional site is created between the two domains. In the former case, the order of the domains is not important, whereas in the latter case, it clearly is.

**Figure 10** Domain compositions of proteins all possessing the supradomain pair (*shaded*), comprising the ATP grasp domains that are involved in binding ATP.
where extensive residue insertions cause structural embellishments to the common core of the domain (86, 90, 91).

Functional properties of enzymes have been very well characterized, and the effect of structural changes on function can therefore be more easily explored (86). In about one third of the most functionally promiscuous enzyme families, relatives have undergone large structural changes. Structural embellishments and residue changes frequently modify the geometry and electrostatics of the active site, resulting in changes in the shape and charge of the substrates binding there (86, 90, 91) (Figure 11).

Structural embellishments can also affect surface properties and are associated with changes in interfaces between domain partners or interfaces involved in oligomerization of the protein to form the biological unit. Again, studies of enzyme families have established the extent to which changes in domain partnership can modulate domain function. In 90% of functionally promiscuous families, change in domain partnership was a frequent mechanism for inducing functional change (86), usually by altering the nature and the geometry of the active site. This change usually has an effect on substrate specificity, whereas the chemistry performed within the active site is often conserved or semiconserved (86).

Figure 11 The ATP-dependent carboxylase-amine/thiol ligase superfAMILY is large and structurally and functionally diverse. All relatives comprise three domains and bind Mg\(^{2+}\)ATP in a cleft between a large and small domain. Substrate specificity varies between relatives, and the enzymes act on a vast array of donor and acceptor substrates. However, the chemistry is conserved throughout involving the ATP-dependent ligation of a substrate carboxylate to an amine or thiol group of a second substrate. In all relatives, the substrate-binding site is located on the large domain. However, some relatives (e.g., D-alanine D-alanine ligase) contain a large structural embellishment, comprising an additional \(\beta\)-sheet that encompasses the active site enclosing it in a box-like conformation that severely restricts access.
(Figure 11). However, even in those families where function is poorly conserved, relatives often possess a common chemical intermediate, although chemical routes to and from this intermediate may vary.

The most highly populated, sequence diverse families also tend to have the most domain partners (1, 57). For most domain partnerships, domains are only found in one combination, AB or BA (93). This conservation of order suggests that this recombination only occurred once during the course of evolution and is also reflected in the conservation of the interfaces mediating the domain interaction (93).

Domain duplication has clearly become an important evolutionary mechanism because once basic functional modules (e.g., like the nucleotide-binding Rossmann domain) evolved early in life, it was quicker to evolve new functions by slightly modifying these domains than by ab initio invention of new domains. Furthermore, once the DNA repair mechanism evolved, it was less likely that new domains would arise through accumulation of errors (93).

How Does Evolution of New Functions in Domain and Protein Families Influence the Evolution of New Pathways and Processes in an Organism?

The preservation of chemistry within a family influences the manner in which biological pathways have evolved within organisms (94, 95). Examination of metabolic pathways in *Escherichia coli*, for which a great deal of experimental data exist, indicates no significant tendency for sequential reactions along a pathway to be performed by homologues (94), disputing the theory that complex pathways arose by gene duplication, with relatives recruited to sequential steps along the pathway for their ability to bind similar substrates/products (96). The picture that emerges from domain distributions in *E. coli* more closely resembles a patchwork model of pathway evolution (97), whereby proteins are recruited for the chemistry they perform. Thus paralogs can be participants of several different pathways. Clearly, nature has found it easier to engineer new substrate specificities than to modify chemistry (86).

**HOW DOES THE EVOLUTION OF PROTEIN FAMILIES INFLUENCE THE COMPLEXITY AND EVOLUTION OF ORGANISMS?**

The function of a protein operates on a number of levels. At the lowest level we can consider the biochemistry, for example, the substrates bound by the protein and the chemistry performed on these substrates. However, there are also the cellular roles of proteins, their locations, and regulation of their expression and activity. In eukaryotes, alternative splicing and posttranslational processing can have significant effects on the functions of the proteins (98).
At the highest level, the effect of proteins on the various phenotypes should be considered, e.g., contribution to the complexity of the organism and how this affects the organism’s ability to respond to diverse environments or acquire new physiological features. As new genome data emerge and comparative genome analysis illuminates commonalities and differences in the protein family repertoires between species, we can start to consider how evolution of protein families influences the complexity of organisms.

Evolution of Families Associated with Regulation

Genome analyses, particularly those exploiting structural data, have clearly revealed the extent of domain duplication within genomes (99, 100). As genomes become larger, a higher proportion of regulatory genes are needed to control the increased interactions resulting from these expansions and thereby enable the most benefit to be gained from increases in the functional repertoires. Comparisons of the regulatory gene complements in *E. coli* and yeast have shown that the transcription factors belong to a limited repertoire of DNA-binding domain families (99). However, their relative abundance often shows dramatic variation among different phylogenetic groups. Again, power laws are observed in the regulatory networks, whereby a select few transcription factor families participate in the regulation of a disproportionately large number of target genes. Teichmann and coworkers view these hubs as global regulators as opposed to the remaining transcription factors that can be considered as fine-tuners (99, 100).

The dramatic differences in transcription factors in different lineages suggest that there have been massive, lineage-specific expansions of these families (99, 100). In fact, about 45% of regulatory interactions in *E. coli* and yeast arose by duplication with inheritance of interaction. Duplication of transcription factors and target genes followed by acquisition of new interactions also occurred (52% *E. coli*, 43% yeast), with or without loss of the original interaction (99, 100).

Other interactions are created by lineage-specific innovation of transcription factors, domain shuffling, and recombining DNA-binding domains with different sensor and signaling domains. Transcription factors are less conserved than their target genes, suggesting that regulation of genes evolves faster than the genes themselves. Regulatory networks also consist of regulatory motifs and modules. Although it appears that motifs have evolved by duplication of complete sets of interactions, there is no evidence to suggest that motifs are preferentially conserved across organisms (99).

Evolution of Immunoglobulin Domain Superfamily in Eukaryotes

Other interesting insights into protein family expansion and the ensuing contribution to organism complexity have come from studies of the immunoglobulin superfamily in eukaryotes (11). Proteins from this family have important roles in
cell-cell communication, for example, as cell-surface receptors during embryonal development. Again, by studying relatives assigned to structural families, it is possible to identify dramatic changes occurring in different organisms.

Comparisons of relatives identified in the fruitfly and the worm illustrate how differences in the immunoglobulin repertoires of these two organisms have contributed to their physiological differences. In the fruitfly, specific expansions function mainly in cell-cell communication, e.g., as receptors that are involved in neuronal development, whereas in the worm, the relatives from this family are largely muscle or extracellular matrix proteins. The fly-specific expansion appears to contribute to higher physiological complexity in the fruitfly by equipping it, for example, with a more intricate nervous system (11).

In their analysis, Vogel and coworkers draw attention to two types of expansions in protein family repertoires: conservative expansion and progressive expansion. Conservative expansion increases an organism’s ability to adapt to its environment but does not significantly affect physiology. This is seen in the expansion of two chemoreceptor families in the worm but not in the fruitfly. It is also seen in the dramatic expansions in some metabolic families in bacteria, as described below. On the other hand, progressive expansion leads to significant changes in form, that may promote an organism’s ability to adapt and may also contribute to divergence of species.

Evolution of Domain Superfamilies in Bacteria

In considering groups of families rather than individual families, analysis of nearly 100 bacterial genomes showed about 200 domain families common to all species (12). We can consider how expansions in these families influence the complexity of the individual genomes. Some common families, about 27, have clearly undergone no expansion. These are mostly associated with protein biosynthesis, supporting the hypothesis of Woese and others (101) that the functions of such critical proteins would have crystallized very early in evolution and that further change would have been undesirable.

Other common families have clearly expanded in the larger genomes in a variety of ways. About 30 families have expanded linearly and dramatically with increasing genome size. These are mostly associated with metabolism and are the most highly expanded families. A further 27 families have expanded nonlinearly with genome size (Figure 12) and are predominantly involved in regulation (12). Expansions of these two types of families are clearly contributing to increases in complexity, as they dominate the genomes and account for approximately 43% of the structural annotations.

The balance between the metabolic and regulatory repertoires may be a factor influencing the restriction on genome sizes observed in bacteria (13). Bacteria have evolved a survival strategy that depends on fast replication of their genomes. Duplication and diversification of metabolic domains significantly broaden the functional repertoires, allowing these more complex bacteria to exploit their
Figure 12 The balance between expansion of metabolic families and regulatory families with increase in genome size. The linear increase in the number of domains primarily involved in metabolism is shown by the thick black line, and the nonlinear increase in the number of domains primarily involved in regulation is shown by the dotted line.

environments more effectively. However, this functional advantage is clearly not without cost since larger genomes are associated with a much greater increase in regulatory genes (see Figure 12). The optimal genome size for bacteria may therefore correspond to a point where there is maximum expansion in the metabolic repertoire for minimal regulatory cost.

CONCLUSION

We approach an exciting era in evolutionary biology as the complement of completed genomes grows and the tools and techniques for analyzing the data increase in sophistication. Already we can see that relatively few domain families (<100) dominate the genomes, influencing their complexity. Different combinations of these domains have clearly made significant contributions to diversity in functional repertoires, phenotypes, and physiologies between organisms. However, we can also see that there is diverse regulation of these families in different lineages.

Other types of data are needed to fully understand the interplay between these different families and their contribution to organism complexity. The recent revolutions in experimental biology— with transcriptomics and proteomics illuminating the behavior of whole protein networks, rather than individual proteins—will help enormously in determining the ways in which new processes and even species have emerged. Sound strategies for integrating and mining these heterogeneous data will be essential (102). However, at any given time, the genome data, both theoretical and experimental, still provide only a current snapshot of the processes
operating during evolution. We should continuously reevaluate our theories in the light of new data.

We are certainly on the cusp of one of the most exciting eras in biology, when the genome data will illuminate our understanding both of the evolution and the nature of biological processes. It is truly amazing that such incredible biological complexity can arise from a relatively small basis set of domain families, many of which are ancient. Complexity in both lower and higher organisms has evolved through local divergence of these families to perform related but different functions, and through formation of a limited number of domain combinations. Taken together with the evolution of differential expression, this complexity has allowed the emergence of the glorious and awe-inspiring biodiversity that we observe today.

ACKNOWLEDGMENTS

We would like to acknowledge Oliver Redfern, Alastair Grant, David Lee, Annabel Todd, and Gabrielle Reeves for contributing data and figures for this review and in particular thank Oliver Redfern for his help in collating all the data and figures for this review.

The Annual Review of Biochemistry is online at http://biochem.annualreviews.org

LITERATURE CITED

17. Todd A. 2003. In Bioinformatics: Genes, Proteins and Computers, ed. CA Orengo,
58. Bateman A, Coin L, Durbin R, Finn RD,
65. Pearl F, Orengo CA. 2003. See Ref. 17, pp. 102–19
87. Deleted in proof
90. Todd A. 2003. See Ref. 17, pp. 151–74
ORENGO ▪ THORNTON


CONTENTS

FROM PROTEIN SYNTHESIS TO GENETIC INSERTION,
Paul Zamecnik 1

THE BIOCHEMISTRY OF PARKINSON’S DISEASE,
Mark R. Cookson 29

APPLICATIONS OF DNA MICROARRAYS IN BIOLOGY,
Roland B. Stoughton 53

ZONA PELLUCIDA DOMAIN PROTEINS, Luca Jovine, Costel C. Darie,
Eveline S. Litscher, and Paul M. Wassarman 83

PROLINE HYDROXYLATION AND GENE EXPRESSION,
William G. Kaelin Jr. 115

STRUCTURAL INSIGHTS INTO TRANSLATIONAL FIDELITY,
James M. Ogle and V. Ramakrishnan 129

ORIGINS OF THE GENETIC CODE: THE ESCAPED TRIPLET THEORY,
Michael Yarus, J. Gregory Caporaso, and Rob Knight 179

AN ABUNDANCE OF RNA REGULATORS, Gisela Storz, Shoshy Altuvia,
and Karen M. Wassarman 199

MEMBRANE-ASSOCIATED GUANYLATE KINASES REGULATE ADHESION
AND PLASTICITY AT CELL JUNCTIONS, Lars Funke, Srikanth Dakoji,
and David S. Bredt 219

STRUCTURE, FUNCTION, AND FORMATION OF BIOLOGICAL
IRON-SULFUR CLUSTERS, Deborah C. Johnson, Dennis R. Dean,
Archer D. Smith, and Michael K. Johnson 247

CELLULAR DNA REPLICASES: COMPONENTS AND DYNAMICS AT THE
REPLICATION FORK, Aaron Johnson and Mike O’Donnell 283

EUkARYOTIC TRANSLESION SYNTHESIS DNA POLYMERASES:
SPECIFICITY OF STRUCTURE AND FUNCTION, Satya Prakash,
Robert E. Johnson, and Louise Prakash 317

NOD-LRR PROTEINS: ROLE IN HOST-MICROBIAL INTERACTIONS AND
INFLAMMATORY DISEASE, Naohiro Inohara, Mathias Chamaillard,
Christine McDonald, and Gabriel Nuñez 355
CONTENTS

REGULATION OF PROTEIN FUNCTION BY GLYCOSAMINOGlyCANS—AS
EXEMPLIFIED BY CHEMOKINES, T.M. Handel, Z. Johnson, S.E. Crown,
E.K. Lau, M. Sweeney, and A.E. Proudfoot 385

STRUCTURE AND FUNCTION OF FATTY ACID AMIDE HYDROLASE,
Michele K. McKinney and Benjamin F. Cravatt 411

NONTEMPLATE-DEPENDENT POLYMERIZATION PROCESSES:
POLYHYDROXYALKANOATE SYNTHASES AS A PARADIGM,
JoAnne Stubbe, Jiamin Tian, Aimin He, Anthony J. Sinskey,
Adam G. Lawrence, and Pinghua Liu 433

EUKARYOTIC CYTOSINE METHYLTRANSFERASES, Mary Grace Goll
and Timothy H. Bestor 481

MONITORING ENERGY BALANCE: METABOLITES OF FATTY ACID
SYNTHESIS AS HYPOthalAMIC SENSORS, Paul Dowell, Zhiyuan Hu,
and M. Daniel Lane 515

STRUCTURE AND PHYSIOLOGIC FUNCTION OF THE LOW-DENSITY
LIPOPROTEIN RECEPTOR, Hyesung Jeon and Stephen C. Blacklow

COPPER-ZINC SUPEROXIDE DismUTASE AND AMYOTROPHIC LATERAL
SClEROSIS, Joan Selverstone Valentine, Peter A. Doucette,
and Soshanna Zittin Potter 563

THE STRUCTURE AND FUNCTION OF SMC AND KLEISIN COMPLEXES,
Kim Nasmyth and Christian H. Haering 595

ANTIBIOTICS TARGETING RIBOSOMES: RESISTANCE, SELECTIVITY,
SYNERGISM, AND CELLULAR REGULATION, Ada Yonath 649

DNA MISMATCH REPAIR, Thomas A. Kunkel and Dorothy A. Erie

GENE THERAPY: TWENTY-FIRST CENTURY MEDICINE, Inder M. Verma
and Matthew D. Weitzman 681

THE MAMMALIAN UNFOLDED PROTEIN RESPONSE, Martin Schröder
and Randal J. Kaufman 711

THE STRUCTURAL BIOLOGY OF TYPE II FATTY ACID
BIOSYNTHESIS, Stephen W. White, Jie Zheng, Yong-Mei Zhang,
and Charles O. Rock 791

STRUCTURAL STUDIES BY ELECTRON TOMOGRAPHY: FROM CELLS
to MOLECULES, Vladan Lučić, Friedrich Förster,
and Wolfgang Baumeister 833

PROTEIN FAMILIES AND THEIR EVOLUTION—A STRUCTURAL
PERSPECTIVE, Christine A. Orengo and Janet M. Thornton 867