EVOLUTION OF DOMAIN FAMILIES

By CHRIS P. PONTING,* JÖRG SCHULTZ,† RICHARD R. COPLEY,† MIGUEL A. ANDRADE,† and PEER BORK†

*National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, Maryland and †EMBL, Meyerhofstrasse, Heidelberg, Germany 20894

I. Introduction

The use of sequence information to frame structural, functional, and evolutionary hypotheses represents a major challenge for the postgenomic era. Central to an understanding of the evolution of sequence families is the concept of the domain: a structurally conserved, genetically mobile unit. When viewed at the three-dimensional level of protein structure, a domain is a compact arrangement of secondary structures connected by "linker" polypeptides. It usually folds independently and possesses a relatively hydrophobic core (Janin and Chothia, 1985). The importance of domains is that they cannot be divided into smaller units—they represent a fundamental building block that can be used to understand the evolution of proteins.

Experience gained from protein structure determination in the past 30 years demonstrates that domains possessing similar sequences also possess similar folds, leading to the inference that such domains are members of homologous families (Doolittle, 1995; Henikoff *et al.*, 1997). Some homologous domain sequences have diverged considerably beyond the level at which homology can be reliably predicted. However,

ADVANCES IN PROTEIN CHEMISTRY, Vol. 54

0065-3233/00 $30.00

from the tertiary structures of these domains it is often seen that their folds and some structural characteristics are conserved even when their sequences are not (Murzin, 1998).

It has been suggested that evolution has generated only approximately 1000 structurally distinct domains (Chothia, 1992; Green *et al.*, 1993). Consequently, the emergence of novel functions during evolution appears to have been more often the result of gene duplication than *de novo* creation of genes from accumulated mutations of noncoding sequence (Ohno, 1970). This often enables us to trace the evolutionary history of proteins and thus make inferences about their functional properties.

This chapter anticipates the completion of *Arabidopsis thaliana, Drosophila melanogaster,* and *Homo sapiens* genome sequencing projects by reviewing current ideas of the evolution of sequence families. In parallel the related issue of domain homolog detection is discussed in light of continuing efforts to map the complete set of domain families.

A. *Protein Annotation*

1. Detection of Sequence Families

Detection of domain homologs in sequence databases depends on their sharing considerable sequence similarities. Although methods such as FASTA or BLAST, which search a single sequence against a database, will detect clearly related homologs, it is estimated that only approximately one-third of all homologs are detectable by such methods (Park *et al.*, 1998). Sensitivity can be improved using initially detected homologs as starting points for further database searches (Park *et al.*, 1997), and this procedure can be iterated for still better detection of diverse homologs (Salamov *et al.*, 1999). For a review of database search methods, see the chapter by Bateman and Birney, in this volume.

To capture the sequence diversity and the conserved features of a protein family, it is necessary to build a multiple sequence alignment using a program such as ClustalW (Thompson *et al.*, 1994). This program highlights residues that are well conserved within a particular protein family, and hence those to which a greater weight should be given when searching a database. Constructing alignments can be a time-consuming procedure. The PSI-BLAST program from the NCBI (Altschul *et al.*, 1997; Altschul and Koonin, 1998) uses the results of a BLAST database scan to construct an internal alignment of the query and database sequences. This alignment is then used to construct a set of position specific scores, termed a profile, used for a subsequent round of database searching, and from this, a new alignment and profile are constructed. The procedure is repeated until either no new sequences are found, or a specified number of iterations is performed. The method is fast and easy to use and has the potential to detect many more homologs than single sequence methods.

A more systematic exploitation of the information available in a multiple sequence alignment is provided by hidden Markov model (HMM)-based methods (Krogh *et al.*, 1994). These methods construct a probabilistic model of a multiple sequence alignment, including statistical weights for which types of amino acid are found at particular positions in the alignment, and information on the regions of the alignment for which there is a high probability of a gap (insertion/deletion) position. Using appropriate software (http://hmmer.wustl.edu/; http://www.cse.ucsc.edu/research/compbio/sam.html), a HMM can be constructed from a preexisting alignment, and used to search a database of protein sequences. Although HMM-based methods are considerably slower than PSI-BLAST, and the construction of the alignment labor intensive, they appear to offer improved detection of homologs (Park *et al.*, 1998). The formalism for profile-based and HMM-based methods are equivalent (Bucher *et al.*, 1996).

The multiple sequence alignments used in profile or HMM construction must span either the entire length of single domains or repeats or domain/repeat combinations that are always found together. Searches employing alignments that encompass multiple domains that are otherwise found in separate proteins result in erroneous annotation of homologs (Bork and Koonin, 1998). In addition, searches employing alignments that encompass multiple repeats result in inaccurate prediction of repeat numbers.

Construction of multiple alignments of homologs using automated methods including PSI-BLAST (Altschul *et al.*, 1997), HMMER (S. Eddy, unpublished), and the Clustal suite of programs (Thompson *et al.*, 1994) are widely acknowledged to produce useful, yet suboptimal, alignments. Ideally, a set of multiple alignments constructed from three-dimensional structures of the homologs would provide the basis for complete detection of all members of homologous families. Determined structures, however, currently represent only a small proportion of sequence space, even when close homologs are considered. Before completion of projected structural genomics programs (Shapiro and Lima, 1998) it is more fruitful to manually optimize alignments to specifications discussed elsewhere (Bork and Gibson, 1996).

Several groups have separately embarked upon projects to generate hand curated gapped multiple alignment libraries for use in homolog detection. The largest collection is Pfam-A (v5.0) (Bateman *et al.*, 1999), which contains 2008 alignments relating to domains and repeats of diverse cellular functions. The SMART library, (v3.0) (Schultz *et al.*, 1998; Ponting *et al.*, 1999a) has a different focus, containing 450 alignments mostly representing genetically mobile domains and repeats with intracellular and extracellular signaling functions. In addition, several other useful facilities are based not on gapped multiple alignments, but on profiles (PROSITE: Hofmann *et al.*, 1999), ungapped alignment blocks (BLOCKS: Henikoff *et al.*, 1999; PRINTS: Attwood *et al.*, 1999), or collections of proposed orthologs (COGS: Tatusov *et al.*, 1997; Koonin *et al.*, 1998). All of these methods allow WWW-based searches of user-supplied sequences against the libraries and thus provide an invaluable complement to the more familiar gapped BLAST searches (Altschul *et al.*, 1997; Hofmann, 1998).

2. *Problems in Protein Annotation*

Database searching using the algorithms just described can be used for the reliable identification of homologs in sequence databases. The value of inferring homology is that it enables the possibility of accurately transferring functional information from the database sequence to the query sequence. Each stage of such analyses is fraught with complications. For example, the sequence itself may be incorrect, inevitably leading to incorrect annotation or a correct sequence may be incorrectly annotated. The varieties of problems related to functional annotation, are discussed next.

a. Interpretation of Genomic Sequence. Incorrect interpretation of the genomic (i.e., the DNA) sequence is one of the main sources of error in protein annotation (Fig. 1, see Color insert). Even in prokaryotic genomes, which contain higher gene densities and simpler gene structures than those of eukaryotes, it is relatively common for detailed analysis to reveal open reading frames (ORFs) that remain unannotated, thereby excluding them from protein databases. Moreover, it can be difficult to find the exact start and stop codons of a particular gene, thus leading to artificial truncation or elongation in the corresponding sequence database entry. Frameshifts represent another potential source of artificial truncation of the protein translations of DNA sequences.

Given the limited accuracy of gene prediction algorithms, it is likely that there will be numerous examples of missed genes in the intron-rich genomes of most eukaryotes. An even more frequent problem than missing ORFs is genes that have not been properly translated. This may mean that introns have been translated, or exons have been missed, although it is important to note that many genes have several alternative splice variants, so that although any one particular translation may be correct, the complete picture of a gene structure may be incomplete. In other cases, single genes may be falsely represented as two protein products, or independent genes may be artificially fused in the process of annotating the genomic data.

These problems lead to the conclusion that the original genomic sequence data should be used as a reference when studying a particular protein of interest, especially when it appears that the standard protein translation of that sequence is in conflict with expectations.

b. *Functional Inference from Homology.* A second source of problems in protein annotation occurs in the process of functional transference between protein sequences related by similarity (Fig. 2, see Color insert). Numerous problems exist in function annotation (e.g., Bork and Bairoch, 1996; Bork and Koonin, 1998; Andrade *et al.*, 1999a; Smith and Zhang, 1997; Doerks *et al.*, 1998). Problems range from semantics and nomenclature to the difficulty of describing complex functions that operate on different linear scales, such as those relating to residues, domains, molecules, and cells. A few functional features, for example, molecular binding partners, localizations, and disease-related variants, are currently annotated in databases, although often in complex syntactical forms that are difficult to parse automatically. Other features such as RNA and protein expression levels and expression distributions are yet to be exploited (Bork *et al.*, 1998).

It is usually impossible to trace the provenance of database annotations. Consequently, even correct annotations may be difficult to verify. It is notable that for one of the smallest prokaryotic genomes, the annotation of function is fundamentally wrong in at least 8% of the entries (e.g., Brenner, 1999). Furthermore, erroneous annotations have been observed to propagate to newly deposited sequences, owing to the use of methods that automatically transfer functional information between sequences sharing significant sequence similarity (Bork and Koonin, 1998). Similarities to functionally characterized database sequences are often overlooked or else not fully exploited. More problematically, a similarity to a database protein is often overinterpreted in terms of function. For example, an "alcohol dehydrogenase" function might be inferred from the closest hit, although query and database proteins share only a common fold and NADH binding site.

A major problem in function prediction is the multidomain nature of many proteins, where a protein can be assigned the function of another, even though it may only share a single common domain. Such

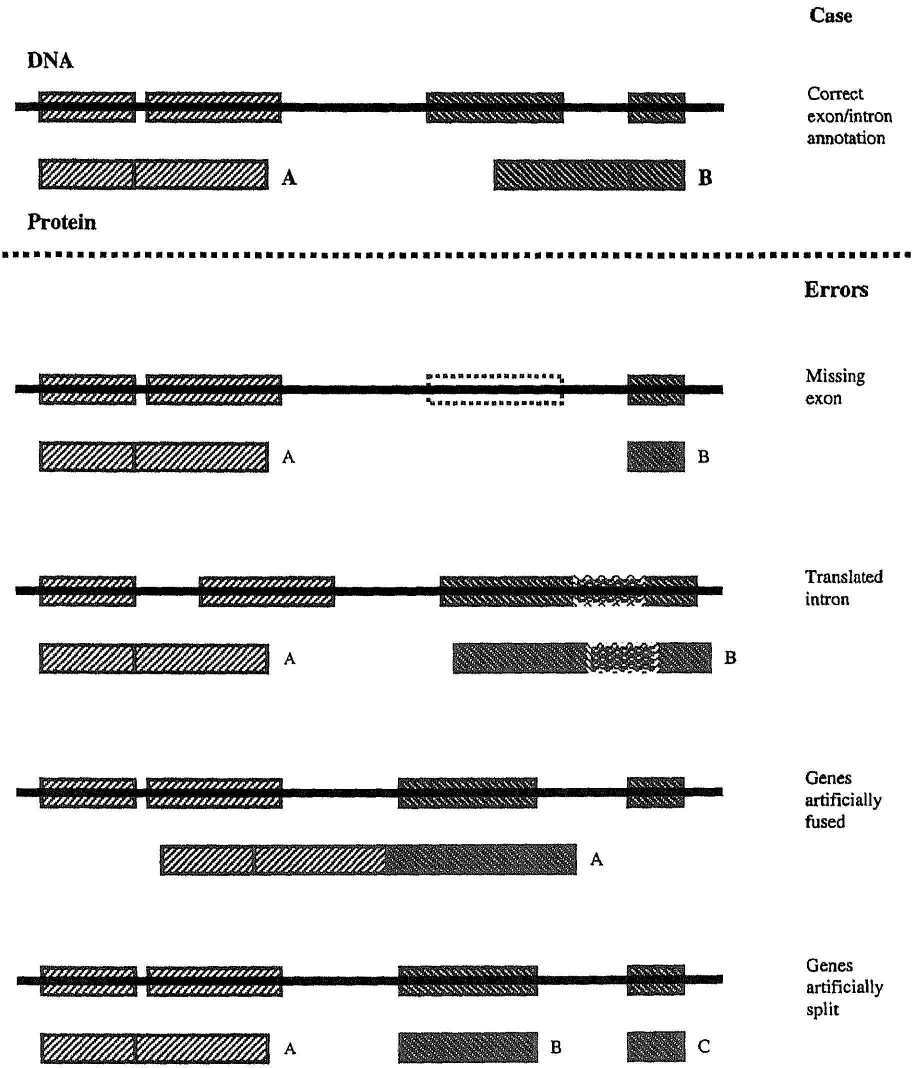
Legends for Color Insert

FIG. 1. Errors arising from the incorrect annotation of protein B from genomic data. The correct annotation lies above the dotted line, and incorrect cases lie below the dotted line. Objects colored in red indicate errors. Similar shading of objects implies homology. Other possible errors that are not represented are the incorrect interpretation of single nucleotide polymorphisms (SNPs) of a gene as different genes, and incomplete detection of splice variants.

FIG. 2. Functional annotation of protein B from sequence similarities to protein A. The correct annotation lies above the dotted line. Incorrect cases lie below the dotted line. Objects colored in red indicate errors in annotation. Similar shading of objects implies homology.

FIG. 3. Three-dimensional structures of three examples of superstructures formed by sequence repeats: a linear rod (the spectrin α-chain dimer [PDB:2spc]), a superhelix of repeats (armadillo repeats of importin α-subunit [PDB:1bk5]), and a closed β-propeller (WD40 repeats from a fragment of the β-subunit of the guanine nucleotide binding protein 1 [PDB:1gg2 chain B]).

FIG. 4. Multiple alignment of the putative protein 4.1-binding motif in syndecans ("SDC") and neurexins ("Neur"). The neurexins are a family of receptors that provide the link between the extracellular environment and intracellular signaling pathways (Littleton *et al.*, 1997; Missler and Südhof, 1998). PDZ domain-containing proteins are known to bind the neurexins and glycophorin via their C-terminal EY[Y/F][I/V] sequences (Littleton *et al.*, 1997). The sequence intervening between the membrane spanning segment and the PDZ domain-binding motif of neurexins and glycophorin contains a protein 4.1-binding motif (Marfatia *et al.*, 1995; Littleton *et al.*, 1997). This motif was found additionally in all known syndecans that function in growth factor signaling and cell adhesion (Rapraeger and Ott, 1998; Zimmermann and David, 1999). The similarity between the neurexin and syndecan families extends beyond this sequence similarity, since their proposed protein 4.1-binding motifs (**4.1m**) both lie on the cytoplasmic side juxtaposed to the transmembrane sequence. Consequently, syndecans are predicted to be protein 4.1-binding proteins. This would be consistent with the known colocalization of syndecan-2 and protein 4.1 at the basolateral membrane of epithelial cells (Cohen *et al.*, 1998). Residues are colored according to an 80% consensus calculated using http://www.bork.embl-heidelberg.de/Alignment/consensus.html; N. Brown and J. Lai, unpubl.): big ("b") residues (E,F,I,K,L,M,Q,R,W,Y) are highlighted in gray, hydrophobic ("h") residues (A,C,F,I,L,M,V,W,Y), aromatic ("a") residues (F,H,W,Y) and aliphatic ("l") residues (I,L,V) are shaded in yellow, charged ("c") residues (D,E,H,K,R) and positively charged ("+") residues (H,K,R) are shown in red, polar ("p") residues (D,E,H, K,N,Q,R,S,T) are shown in brown, and small ("s") residues (A,C,S,T,D,N,V,G,P) and tiny ("u") residues (A,G,S) are shown in green. GenBank identifier (gi) accession codes and residue limits are shown following the alignment. Predicted secondary structure (Rost and Sander, 1993) is shown beneath the alignment (h/H represents helix and e/E represents β-strand); expected accuracies are greater than 82% (upper case) or greater than 72% (lower case). CIOSA, *Ciona savignyi;* DROME, *Drosophila melanogaster,* HUMAN, *Homo sapiens;* MOUSE, *Mus musculus;* RAT, *Rattus norvegicus.*



Case
DNA
Correct exon/intron annotation
A
B
Protein
Errors
Missing exon
A
B
Translated intron
A
B
Genes artificially fused
A
Genes artificially split
A
B
C

FIG. 1

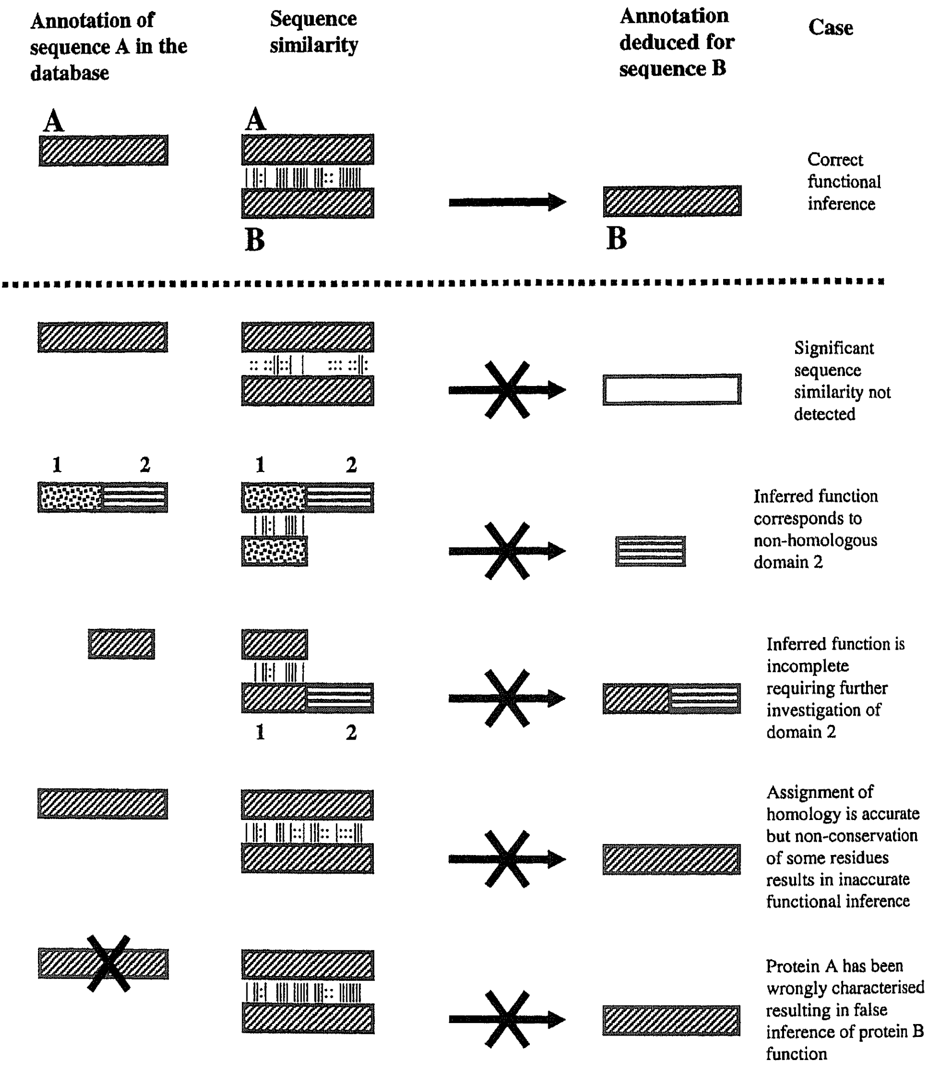

Annotation of sequence A in the database
Sequence similarity
Annotation deduced for sequence B
Case
A
A
B
B
Correct functional inference
Significant sequence similarity not detected
1
2
1
2
Inferred function corresponds to non-homologous domain 2
Inferred function is incomplete requiring further investigation of domain 2
1
2
Assignment of homology is accurate but non-conservation of some residues results in inaccurate functional inference
Protein A has been wrongly characterised resulting in false inference of protein B function

FIG. 2

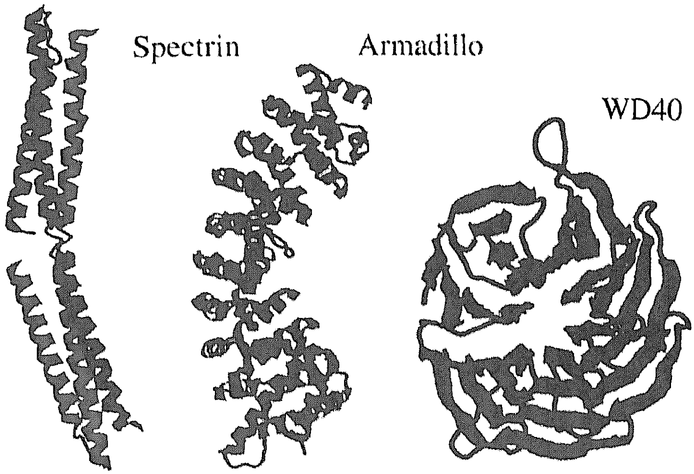


FIG. 3

BL1A/HUMAN	LGHYLIRHKGTYLTHEAKG	not yet	387- 405
IGSF4/HUMAN	LGRYFARHKGTYFTHEAKG	4519602	395- 413
GlycoC/HUMAN	MLRYMYRHKGTYHTNEAKG	183327	78- 96
F22162_1/HUMAN	MVWCSVRQKGSYLTHEASG	3451335	337- 355
NeurIV/DROME	IGRYLHRHKGDYLTHEDQG	1518221	1237-1255
NeurIV/MOUSE	FYLQNHRYQGSYHTNEPKA	2773268	1305-1323
NeurI/RAT	MYKYRNRDEGSYHVDESRN	1083730	413- 431
NeurIII/RAT	MYKYRNRDEGSYQVDETRN	539981	1417-1435
SDC3_CHICK	IYRMKKKDEGSYTLEEPKQ	1351050	371- 389
SDC/CIOSA	AYRMRKKDEGSYALDEKKP	4519942	177- 195
Consensus	bh+bb.RccGoYhscEs+s		
2-structure	eee e		

FIG. 4

```
T12D8.4/CAEEL        ILFNNPEQAYFAGQEISGKVIIENKEPKKVNEILLELKG 39 KDGKEKERILPAGIHQVPFSYTLP    9-110 3879793
ZK938.2/CAEEL        ISFDKQSEPYYPGDTVKGTVVLTNKTPLDARCVTIKCRG 25 WVSKDGQNKMPEGGYIWTFEIKLP   10- 97 3881905
Y51B9A.4/CAEEL       IVFSNPSKTYLPGDYVSGKVLLTTKDPISARYMEITWKG 24 WIAKDGVDTIPAGTLKSRFRFRLP    5- 91 3881067
Y49E10.24/CAEEL      LGLSNCEQIYEPGGTIEGYVTFDLRESVKVKAIRISAEG 39 PSEGHSKSSVFPGIHVYPFKFQIP    7-108 3979997
R06B9.4/CAEEL        PQISFEKEVYFPGDEVKGRAWVSTTKNLKATSVEITFSG 31 WTPEDTENTFPSGDYEWNFSFELP    8-101 3878866
R06B9.2/CAEEL        VLFDKPNAVYAAGQKISGRVVFSTASQQNPRWIDVQLHG 43 WRKTDKAARLLPGKYEWQFWFQLP    6-111 3878867
VDUP1/HUMAN          VVFNDPEKVYGSGEKVAGRVIVEVCEVTRVKAVRILACG 32 TGENEMVIMRPGNKYEYKFGFELP   12-106  688297
F40F8.8/CAEEL        FDIRLNKDVYYAGETISGSVLLENTENIKIRGIRVLLRG 35 SDESDSVPILARGVHQFSFNFDLP    7-104 3876948
palF/EMENI           IEPDDPWRSYFPGDVIKGTVSLTVVRPVRITHLVISLHG 43 EVVLCGEGRLKEGIYKFRFEMSFP   51-156 1755051
M176.1/CAEEL         IVLDRDTCKYRPEECVTGNVVIINRKELKARTLRVYIKG 46 WNFTSDTDCIKPGTHKFPFSFKLP    6-114 3878746
YOW1_CAEEL           INIVLAEPRCMAGEFFNAKVLLDSSDPDTVVHSFCAEIK 26 VQLCDSGTCLPVGKHQFPVQIRIP    6- 94  267456
YGE6_YEAST           IDIDEPHRVWKPNESITGEAVIDIKRDITNVAIKLSLVC 45 VDKTTILNGLSKGEHRFPFRIRIP   41-148 1723826
YA4C_SCHPO           LIVIQGLADSAEPTPLSGSVVLCLNEAISVKAISLKLVG 56 VPYTGANRTLRAGNYEYPFHVMLP   24-119 1175406
C04C11.2/CAEEL       HLDAGDEPTFKGGELITGKLLIELRKPIIINAVKLQMKG 31 KPEPGHFQWIGDFTYSLPFECPLP   14-107 3873953
DCRA_MOUSE           IKIKRANKVYHAGEMLSGVVVISSKDSVQHQGVSLTMEG 29 TIDVLKPGKIPSGKTEVPFEFPLL    7- 98 3334162
YDA5_SCHPO           PLVMYGNPDVSSGALASGLAKLTVFPADLKCESIVMEFV 25 WEFSKEPTVYTHGTHTWPFTFIIP   28-115 1723510
T07F12.3/CAEEL       LELSRKDACYNWGDTIQGKLNLKISEGSIEITSLRILFH 25 SNAISQPIVVSQQEYSIPIEFTLP   75-162  861404
YL92_YEAST           LNPPYNGEFYSSNDQMSGIVSLQLTKALSIRKISVILKG 48 LDGSSKPFKVKPGSYNYSFQFDKF    9-119 2506770
F58G1.6/CAEEL        GTFSTDIFFEKFPIRMKQKRAIRLLVPDRPLRAGCGVCR 25 LPTTKGIDVIPEGHHRLPIFLNVP    6- 93 3877916
ARRH_LOCMI           LEVTLDREIYYHGEKLAANVIINNNSRKTVKNIKVYVVQ 16 SLETREGCPITPGASFTKVFYLVP  206-284  416656
ARRS_BOVIN:1AYR      LAVSLSKEIYYHGEPIPVTVAVTNSTEKTVKKIKVLVEQ 15 VAAEEAQEKVPPNSSLTKTLTLVP  205-267 4139915
spo0M/BACSU          GEEVQGTVHVKGGKIAQDIRYIDLQLSTRYVIVKDDEEH  8 FRVTGSFTIQPGEEHQFPFTFTLP   30-100 3522984
Consensus (80%)      h.hs.sp..b..sp.hpu.shlp.pps.p...h.l.h.h    ....p....b..s.hph.b.b.lP
2-structure          EEEEE  EEE  EEEEEEEEE   EEEEEEEEE         EEEEEE     EEEEEEEEE
```

FIG. 5

```
KIAA0462_HUMAN   LQCPRCSAS----VPANPGVCGNCGEN-VYQCHKCRSINYDEKDPFLCNACGF 3413886  751- 798
PUSH_DROME       LQCPRCSAA----VPAYPGVCGNCGEN-VFQCHKCRAINYDEKDPFLCHSCGF 4426611 3791-3838
C44E4-1A_CAEEL   LHCPRCTVP----VRTHSGVCESCGEN-AFQCAKCRAINYVEKEPFLCQSCGF 2088725 2442-2489
Y45A_METJA       YVCISCNAEIAPREKSTKFPCPNCGEVEIVRCERCRKLNN----PYKCPKCGF 2826285    6-  54
AF0573_ARCFU     TRCVSCGAVLV-GANYVAFPCPECGEM-IYRCKKCRRLSN----PYVCESCGF 2650062    5-  51
consensus/80%    bpCspCsAs....s.shshsCssCGEs.labCcKCR.lN.....PalCpuCGF
2-structure           e                        hhhhhh
```

FIG. 6

WASP_MOUSE	GRGALLDQIR----QGIQLNKT	2499130	448- 465
C07G1.4/CAEEL/2	ARGDVMAQIR----QGAQLKHV	1326381	868- 885
NWASP/BOVINE/1	SKAALLDQIR----EGAQLKKV	1644232	405- 422
NWASP/RAT/2	GRDALLDQIR----QGIQLKSV	2274845	429- 446
CR16/RAT	GRSALLADIQ----QGTRLRKV	4096360	45- 62
C07G1.4/CAEEL/1	GRSNLLAEIQ----AGKQLRSV	1326381	837- 854
R144.4/CAEEL	ARNALLGDIH----KGLKLKKT	746496	28- 45
ProRich/XENLA	GRNALLGDIC----KGAKLKKT	345604	36- 53
VRP1_YEAST	GRDALLGDIR----KGMKLKKA	2507155	30- 47
ACTO_ENTHI	DRNELLSGIK----EGKELKKA	3912973	35- 52
KIAA0429/HUMAN	QGEDMLNAIR----RGVKLKKT	2887433	328- 345
T24B8.4/CAEEL/1	NENAQL-EIK----KGFKLRPT	3880148	25- 41
T24B8.4/CAEEL/2	DRGEFLKGIQ----GGFKLKKT	3880148	173- 190
YAV1_SCHPO	DRSALLQQIH----TGTRLKKT	1723244	1748-1765
VP61_NPVOP/1	NRSALLDQIK----QGKTLKKT	3915911	354- 371
VP61_NPVOP/2	PRSTLLSEIR----QGKTLKKL	3915911	382- 399
ORF1629/HANPV	PRTELMEQIQ----KGIKLKKV	2072251	196- 213
R06C1.3/CAEEL	ARSDLLAQIQ----SGIKLKKV	3878849	442- 459
Y269_HUMAN	ARSVLLEAIR----KGIQLRKV	4507913	497- 514
dJ393P12.2/HUMAN	AHSDLLSAIC----QGFQLRRV	3123552	435- 452
LA17_YEAST	GRDALLASIRG-AGGIGALRKV	2498506	547- 567
KIAA0633/HUMAN/1	LHSALMEAIHS-AGGKDRLRKT	3327080	1164-1184
KIAA0633/HUMAN/2	ERSALLAAIRG-HSGTCSLRKV	3327080	1204-1224
KIAA0633/HUMAN/3	ARQALMDAIRS-GTGAARLRKV	3327080	1292-1312
pp78-81/LDMNPV/1	PTDALFAEIR----RGVQLKPA	3822236	302- 319
pp78-81/LDMNPV/2	TPDALFAEIR----QGVKLKPA	3822236	327- 344
pp78-81/LDMNPV/3	SRAPLLLEIEN--RDKIKLKKV	3822236	356- 375
VRP1_YEAST	MGAPQLGDILA--GGIPKLKHI	2507155	87- 106
orf1035/OPMNPV	ARNLLLEQIK----QKPSLRPV	1903309	264- 281
AA64_HUMAN	SRDQLLAAIR--SSNLKQLKKV	231475	546- 565
Espin/Rat	DNSELLAEIK----AGKSLKPT	3818569	26- 43
CAP1_HUMAN	SRSALFAQINqgeSITHALKHV	399184	253- 274
CAP_SCHPO	DMGAVFAEINkGEGITsgLRKV	543928	333- 353
CAP2_HUMAN	SRSALFAQLNqGEAITkgLRHV	729015	259- 280
CAP_DICDI	GLGAVFGELSkGDGVTsgLKKV	1705592	254- 275
Consensus (80%)	s+ssLb..Ip.....s.pL++s		
2-structure (PHD)	hHHHHHHHh		

FIG. 7

ANGIOP3/MOUSE	GSVNFQR-TWEEYKE-------GFGNVAR-----------EHWLG	4512042	295- 369
FIBB/XENLA	GSVGFGR-TWDSYKS-------GFGNIAANGGKGICDMPGEFWLG	1160337	239- 324
NG27/MOUSE	GSVDFNQ-SWEAYKD-------GFGDPQG-----------EFWLG	4050092	190- 262
FICOLIN2/HUMAN	GSVDFYR-DWATYKQ-------GFGSRLG-----------EFWLG	4758348	103- 176
FIBX_MOUSE	GSTNFTR-EWKDYKA-------GFGNLER-----------EFWLG	120159	204- 278
FIBX/HUMAN	GSENFNR-GWKDYEN-------GFGNFVQKHG--------EYWLG	4758372	81- 157
ANGIOP2/MOUSE	GSVDFQR-TWKEYKE-------GFGNPLG-----------EYWLG	2257931	282- 356
HAKATA/HUMAN	GSVDFFR-SWSSYRA-------GFGNQES-----------EFWLG	4504331	91- 164
TENA/MOUSE	GREDFYR-NWKAYAA-------GFGDRRE-----------EFWLG	91806	1800-1875
FIBA_HUMAN	GSLNFNR-TWQDYKR-------GFGSLNDEGEG-------EFWLG	1706799	630- 710
ANGIOP4/HUMAN	GTVNFQR-NWKDYKQ-------GFGDPAG-----------EHWLG	4512044	289- 363
FIBB_PETMA	GSSNFAR-DWNTYKA-------EFGNIAFGNGKSICNIPGEYWLG	120126	192- 277
FIBH_HUMAN	GSVDFKK-NWIQYKE-------GFGHLSPT-------GTTEFWLG	71828	177- 255
FIBA_PARPA	GTINFYR-SWSYYQT-------GFGNLNT-----------EFWLG	120092	68- 144
TENA/HUMAN	GQLDFFK-RWRSYVE-------GFGDPMK-----------EFWLG	3954838	361- 436
MFA4_HUMAN	GSVSFFR-GWNDYKL-------GFGRADG-----------EYWLG	2506403	39- 113
ANGIOP3/HUMAN	GSVNFFR-NWENYKK-------GFGNIDG-----------EYWLG	4757752	278- 352
FIB2_PETMA	GSLNFNR-SFSAYRE-------GFGTVDGSGHG-------ELWLG	462084	402- 482
CDT6/HUMAN	GLVSFYR-DWKQYKQ-------GFGSIRG-----------DFWLG	2765527	129- 205
TENX/MOUSE	GQTDFWR-DWEEYAH-------GFGNISG-----------EFWLG	2564958	3790-3865
RESTR/CHICK	GLTDFFR-KWADYRV-------GFGNLED-----------EFWLG	86419	1131-1206
FIBG_PETMA	GSVNFTR-DWVSYRE-------GFGYLAPT-------LTTEFWLG	120143	176- 253
D1009.3/CAEEL	GDGSFHRGTMKKFVE-------GFGNLQG-----------SHWLG	1072170	220- 291
SCA_DROME	GSADFNR-SWADYAQ-------GFGAPGG-----------EFWIG	134288	515- 582
FRP3/BIOGL	GNVDFYR-GWKEYRD-------GFGDYNI----------GEFYLG	2317872	186- 254
T01D3.6B/CAEEL	ADLNTNK-TFQDYLI-------GFGNPATQ----------SVWLG	3876747	670- 751
CDD/CHICK	STEITWKESWTTYKY-------GFGDVQG-----------DHWLG	3746539	41- 113
CA25_HUMAN	-SSVPRK-TWWASKS-PDNKPVWYGLDMN--------RGSQFAYG	4502959	1291-1372
CA12/AREMA	-GQIFKG-VWYRGEP----GHVWFADEME--------NGLLVHLQ	1778210	489- 567
CA1B_HUMAN	SEGVRIS-SWPKEKP-----GSWFSEFKR---------GKLLSYL	4502939	1605-1684
CA12/HALDI	-TEYRRD-RWTKDST---SGQYFMSDVFG--------KMKEFKYD	4519617	1185-1264
CA12/PARLI	-SQIINS-TWYVGKV-----KRTYFSTME--------GGDKFSYI	280636	546- 623
CA12/ALVPO	-SQVFKG-SWYSGPQ----KYVWFGEDMD--------NGFQFTYK	5174770	695- 773
CAF1_EPHMU	-NAGDLK-SWSGHSI-------WFSDMLG---------GFKLTYD	115436	371- 447
CAF1/STRPU	-DEISRA-RWYEGAS-----GSRYITEMG---------LEKFSYE	283750	3000-3080
CA12/STRPU	-NVVSNM-TWYVGKT----KRAFFSSMHG---------GDKFAYI	283748	1230-1307
CA11_CHICK	-ATIAQK-NWYLSKNPKEKKHVWFGETMS--------DGFQFEYG	115268	1246-1328
CA11/DANRE	-PKIPRK-NWWTSKS-KAQKHVWFGESMN--------GGFHFSYA	773661	258- 339
CA21_CHICK	-EDIPTK-TWYVSKNPKDKKHIWFGETIN--------GGTQFEYN	115338	1155-1237
Consensus/80%	.s.sh.+.sW..h.........hFGp..............cahbs		
2-structure:1FZC	HHHHHH EEE EEEE		

FIG. 8a

```
ANGIOP3/MOUSE     QDCAEIKR---SGVNTSGVYTIYETNMTK------PLKVFCDMET-DGGGWTLIQHRED
FIBB/XENLA        KECEEIYR---KGGETSEMYLIQPDSFFR------PFKVYCDMAT-HDGGWTVIQNRQD
NG27/MOUSE        RDCQELFQ---EGERHSGLFQIQPLGSP-------PFLVNCEMTS--DGGWTVIQRRLN
FICOLIN2/HUMAN    RTCKDLLD---RGHFLSGWHTIYLPDCR-------PLTVLCDMDT-DGGGWTVFQRRVD
FIBX_MOUSE        KDCSDHYV---LGRRSSGAYRVTPDHRNS------SFEVYCDMET-MGGGWTVLQARLD
FIBX/HUMAN        ADCSEIFN---DGYKLSGFYKIKPLQSPA------EFSVYCDMSD--GGGWTVIQRRSD
ANGIOP2/MOUSE     RDCAEIFK---SGLTTSGIYTLTFPNSTE------EIKAYCDMDV-GGGGWTVIQHRED
HAKATA/HUMAN      RNCRELLS---QGATLSGWYHLCLPEGR-------ALPVFCDMDT-EGGGWLVFQRRQD
TENA/MOUSE        RDCSQAML---NGDTTSGLYTIYINGDKT-----QALEVYCDMTS-DGGGWIVFLRRKN
FIBA_HUMAN        RDCDDVLQ-THPSGTQSGIFNIKLPGSSK------IFSVYCDQET-SLGGWLLIQQRMD
ANGIOP4/HUMAN     QDCAEIQR---SGASASGVYTIQVSNATK------PRKVFCDLQS-SGGRWTLIQRREN
FIBB_PETMA        MHCEDIYR---NGGRTSEAYYIQPDLFSE------PYKVFCDMES-HGGGWTVVQNRVD
FIBH_HUMAN        KDCQDIAN---KGAKQSGLYFIKPLKANQ------QFLVYCEIDG-SGNGWTVFQKRLD
FIBA_PARPA        RDCYDILQSCSGQSPPSGQYYIQPDGGN-------LIKVYCDMET-DEGGWTVFQRRID
TENA/HUMAN        SDCSQVQQ---NSNAASGLYTIYLHGDAS-----RPLQVYCDMET-DGGGWIVFQRRNT
MFA4_HUMAN        LDCDDIYA---QGYQSDGVYLIYPSGPSV------PVPVFCDMTT-EGGKWTVFQKRFN
ANGIOP3/HUMAN     KDCQQAKE---AGHSVSGIYMIKPENSNG------PMQLWCENSL-DPGGWTVIQKRTD
FIB2_PETMA        IDCLDVLQ-RRPGGKASGLYEVRPRGAKR------ALTVHCEQDT-DGGGWTLVQQRED
CDT6/HUMAN        YDCSSLYQ---KNYRISGVYKLPPDDFLGSP----ELEVFCDMET-SGGGWTIIQRRKS
TENX/MOUSE        RDCGEELKN---GPSASKTTTIFLNGNRE-----RPLDVFCDMET-DGGGWLVFQRRMD
RESTR/CHICK       QDCAQHLM---NGDTLSGVYTISINGDLS-----QRVQVFCDMST-DGGGWIVFQRRQN
FIBG_PETMA        KDCQQVVD---NGGKDSGLYYIKPLKAKQ------PFLVFCEIEN--GNGWTVIQHRHD
D1009.3/CAEEL     DNCLERLAL----GSPSGVYSIQSVE---------KFQAFCDMDT-TTGGWTVIQRRVD
SCA_DROME         HDCSEVHT------QTDGLHLIAPAGQRH------PLMTHCTADG-----WTTVQRRFD
FRP3/BIOGL        KSCRDVNS-------TDERVVVTLTS---------GLKVMCDTKT-DGGGWIIFQRRIN
T01D3.6B/CAEEL    RHCADLYV--YWGVRESGVNSINPPFVLPQRAKFAPMNVYCDMTT-NGGGYTLMSSDT-
CDD/CHICK         ADCSRLTS-----SSPSGVYVIQPAQSP-------PRVVWCDMDT-EGKGWTVVQRNTY
CA25_HUMAN        RTCDDLKL--CHSAKQSGEYWIDPNQGSV----EDAIKVYCNMET----GETCISANP-
CA12/AREMA        RTCKDLAM--AHPEFEDGMYWVDPNQGSP----VDAIEVFCDIQA----HQTCVMAKP-
CA1B_HUMAN        RTCKDLQL--SHPDFPDGEYWIDPNQGCS----GDSFKVYCNFTS---GGETCIYPDKK
CA12/HALDI        KNCRDIKL--SNPDFKDGEYWIDPNGDSA----LDALKVFCRMET----LETCIKPKI-
CA12/PARLI        RSCKDIFL--NDANAESGTYWVDPNLGCH----QDAIQVHCEQET----QATCVSPSM-
CA12/ALVPO        RTCKDLAL--AHPEMEDGVYWVDPNQGSP----IDAIEVYCHIKT----HQTCVFAKP-
CAF1_EPHMU        RSCHDLFL--EDNSTSDGYYWIDPNGGCI----GDAVKVFCNFTG--GVQQTCISATK-
CAF1/STRPU        RSCKDLFK--CYPEAEDGNYWIDSNEGSV----KDAFLAHCVKRGESGSPETCITPRV-
CA12/STRPU        RSCKDVFL--NNVEAESGYYWVDPNLGCQ----KDAIQVYCEAET----GATCVPSTN-
CA11_CHICK        RTCRDLKM--CHGDWKSGEYWIDPNQGCN----LDAIKVYCNMET----GETCVYPTQ-
CA11/DANRE        RSCRDLKL--CHPEWKSGDYWVDPNLGSA----ADAIKVFCNMET----GETCVKPST-
CA21_CHICK        RTCRDLRL--SHPEWSSGFYWIDPNQGCT----ADAIRAYCDFAT----GETCIHASL-
consensus/80%     bsCpclb.....s...sG.Y.lp.s.s........sbpVaCcbps....sbTshp.c..
2-structure:1FZC    HHHHHH         EEEEEE             EEEEEE        EEEEEEE
```

FIG. 8b (*continued*)

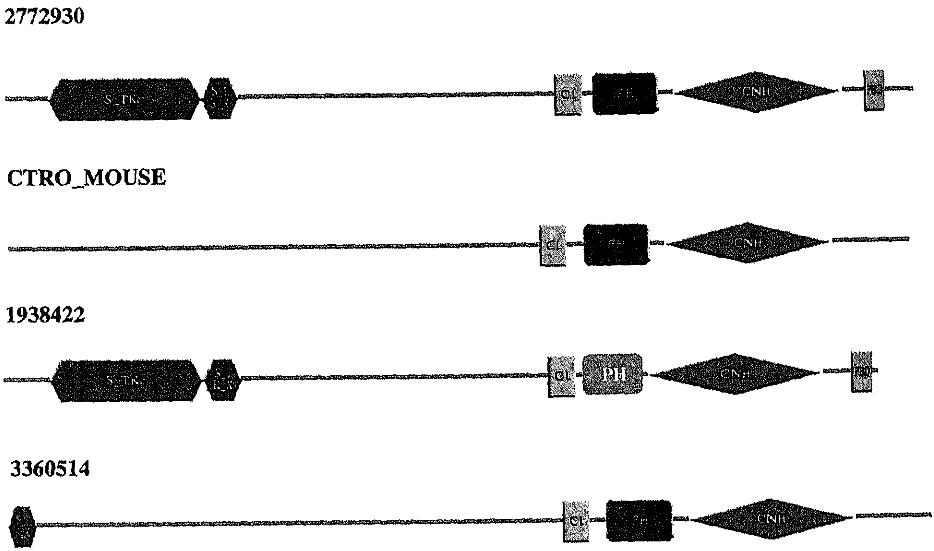

2772930
S_TKc
CL
CNH
CTRO_MOUSE
CL
CNH
1938422
S_TKc
CL
PH
CNH
3360514
CL
CNH

FIG. 9

FIG. 5. Multiple alignment of arrestin homologs prepared using the PROBE algorithm (Neuwald *et al.*, 1997) alignment blocks are shown with numbers in parentheses representing intervening amino acid residues. Significant similarity between these sequences may be shown using PSI-BLAST searches. For example, a search with the sequence of *C. elegans* F48F7.7 demonstrated significant similarities with mammalian arrestins in iteration 1 ($10^{-3} < E < 10^{-12}$), with fungal proteins (such as *S. cerevisiae* Rod1p and YFR022w, and *S. pombe* SPBC24E9.02) in iteration 2 ($10^{-3} < E < 10^{-10}$), and divergent homologs such as *Dictyostelium discoideum* PepA, human Hβ58 and yeast Pep8p in iteration 3($10^{-3} < E < 10^{-10}$) in iteration 3. Also in iteration 2, significant similarity was demonstrated to *Bacillus subtilis* spo0M. The PSI-BLAST algorithm aligned the first of the spo0M two homologous domains with the two domains in eukaryotic homologs with $E = 7 \times 10^{-6}$ and $E = 1 \times 10^{-3}$. The secondary structure shown is that known for bovine S-arrestin (ARRS_BOVIN; PDB: 1AYR). Annotation of this alignment is as Fig. 4. Species abbreviations: as Fig. 4, except: BACSU, *Bacillus subtilis;* CAEEL, *Caenorhabditis elegans;* EMENI, *Emericella nidulans;* LOCMI, *Locusta migratoria;* SCHPO, *Schizosaccharomyces pombe;* YEAST, *Saccharomyces cerevisiae.*

FIG. 6. Multiple alignment of a putative zinc finger in pushover/calossin and two archaeal proteins. Conserved cysteines predicted to bind Zn^{2+} are shown as white-on-black. Annotation of this alignment is as Fig. 4. Species abbreviations: as Figs. 4 and 5, except: ARCFU, *Archaeoglobus fulgidus;* METJA, *Methanococcus jannaschii.*

FIG. 7. Multiple alignment of WH2-like putative actin-binding motifs. Annotation of this alignment is as Fig. 4. Species abbreviations: as Figs. 4 and 5, except: BOVIN, *Bos taurus;* DICDI, *Dictyostelium discoideum;* ENTHI, *Entamoeba histolytica;* XENLA, *Xenopus laevis;* LDMNPV, *Lymantia dispar* multicapsid nuclear polyhedrosis virus; OPMNPV, *Orgyia pseudotsugata* multicapsid polyhedrosis virus; HANPV, *Helicoverpa armigera* nucleopolyhedrovirus; HZNPV, *Helicoverpa zea* nuclear polyhedrosis virus; SLNPV, *Spodoptera littoralis* nuclear polyhedrosis virus.

FIG. 8. Multiple alignment of FBG and COLFI domains. Conserved cysteines predicted to form a single disulphide bridge on the basis of the known tertiary structure of fibrinogen are shown as white-on-black. Annotation of this alignment is as Fig. 4. The secondary structure shown is taken from the known structure of fragment double-D of human fibrin (Spraggon *et al.*, 1997): PDB code IFZC. Species abbreviations: as Figs. 4, 5, and 7, except: ALVPO, *Alvinella pompejana;* AREMA, *Arenicola marina* (lugworm); BIOGL, *Biomphalaria glabrata* (bloodfluke); CHICK, *Gallus gallus;* DANRE, *Danio rerio;* EPHMU, *Ephydatia muelleri* (sponge); HALDI, *Haliotis discus* (abalone); PARLI, *Paracentrotus lividus* (sea urchin); PARPA, *Parastichopus parvimensis;* PETMA, *Petromyzon marinus* (sea lamprey); STRPU, *Strongylocentrotus purpuratus* (sea urchin).

FIG. 9. The domain architectures of proteins with **C1** and **CNH** domains. These are (from the top): *D. melanogaster* Genghis Khan (GenBank identifier [gi]: 2772930), *Mus musculus* citron (SwissPROT: CTRO_MOUSE), *C. elegans* K08B12.5 (gi 1938422) and an alternatively spliced version of mouse citron (gi 3360514). The **PH** domain of K08B12.5 (gi 1938422), depicted in light blue, is not detectable using standard database search procedures. However, similarities in domain architecture and sequence to the other three sequences provides evidence for its occurrence. The mouse citron splice variant (gi 3360514) contains an extension (**S_TK_X**) of a serine/threonine kinase domain (**S_TKc**). However, it lacks the catalytic domain that, in all other cases, is found at its N terminus. As expected, the corresponding sequence is annotated in GenBank as being a fragment, and a catalytic domain is expected in the full-length sequence.

mistakes can be largely avoided using the specialized domain-based search tools described in this chapter.

B. *Domains, Repeats, and Motifs*

This chapter describes the evolution of domain and repeat families that are represented in SMART. The aim in the original version of SMART (Schultz *et al.*, 1998) was to curate multiple alignments of those domains and repeats that are most frequently represented among intracellular signaling proteins of eukaryotes (Table I). Subsequently, SMART was updated to include domains and repeats that are typically seen in eukaryotic extracellular contexts, including extracellular matrix and membrane-bound signaling proteins, and prokaryotic intracellular signaling proteins (Ponting *et al.*, 1999a). The domains and repeats discussed here and represented in SMART appear to be among the most "genetically mobile."

1. *Domain Characterization*

The greater the diversity of a sequence family represented in an alignment, the better the profiles or HMMs derived from it at detecting homologous family members. Thus, considerable care is taken in the construction and updating of the SMART library that *all* detected homologs, assigned using significant statistical estimates for similarity, are represented in an alignment. Consequently, all available sequence similarity algorithms that use rigorous statistical methods are employed in searches for homologs. These include PSI-BLAST (Altschul *et al.*, 1997), MoST (Tatusov *et al.*, 1994), FASTA (Pearson and Lipman, 1988) and HMMER (S. Eddy, unpublished). As database search algorithms have improved considerably since the initial versions of the SMART alignments were constructed, it is appropriate to discuss the protocol now used to distinguish true-positive homologs from false-positive ones.

In the absence of compositionally biased (i.e., those regions that are not typical of globular proteins) sequences, alignment scores resulting in *E* values less than 0.01 (PSI-BLAST), 0.05 (MoST), and 0.1 (FASTA and HMMER2) are considered possible indicators of homology. An *E* or expect-value of an alignment score *X*, is the estimated number of alignments with a score equal to, or greater than, *X* expected from the search purely by chance. However, in most cases more than one method is employed to demonstrate significant sequence similarity. In a small number of cases, similarities with marginal significance are warranted to indicate homology on the basis of orthology (including identical predicted domain architectures) or owing to experimental evidence of

TABLE I

Domain Families Represented in Current Version of SMART[a]

Domain	Definition	Phyletic distribution	Yeast proteins (domains)	Worm proteins (domains)	PDB
14_3_3	14-3-3 homologs	E(MFP)	2(2)	2(2)	1QJA
4.1m	Putative band 4.1 homologs' binding motif	E(M)	0(0)	1(1)	
A4_EXTRA	Amyloid A4	E(M)	0(0)	1(1)	
AAA	ATPases associated with a variety of cellular activities	E(MFP)AB	87(122)	112(159)	1BMF
acidPPc	Acid phosphatase homologs	E(MFP)AB	6(6)	4(4)	
ACTIN	Actin	E(MFP)	10(10)	9(9)	1ALM
ADF	Actin depolymerization factor/cofilin-like domains	E(MFP)	4(5)	4(6)	1CNU
AHL	Domains in archaeal histones and eukaryotic TFs	E(MFP)A	3(3)	4(4)	1A7W
ALBUMIN	Serum albumin	E(M)	0(0)	0(0)	1AO6
alkPPc	Alkaline phosphatase homologs	E(MF)B	1(1)	0(0)	1AJA
ANATO	Anaphylatoxin homologous domain	E(M)	0(0)	0(0)	1C5A
ANK	Ankyrin repeats	E(MFP)B	19(65)	87(462)	1BLX
ANX	Annexin repeats	E(MFP)	0(0)	4(15)	1A8A
AP2	DNA-binding domain in plant proteins such as APETALA2 and EREBPs	E(P)	0(0)	0(0)	1GCC
APPLE	APPLE domain	E(M)	0(0)	1(1)	
ARF	ARF-like small GTPases; ARF, ADP-ribosylation factor	E(MFP)	5(5)	10(10)	1HUR
ARM	Armadillo/β-catenin-like repeats	E(MFP)	2(16)	5(28)	1IAL
AT_hook	DNA-binding domain with preference for A/T-rich regions	E(MFP)AB	5(6)	14(26)	2EZD
ArfGap	Putative GTPase activating proteins for the small GTPase, ARF	E(MFP)	6(6)	8(8)	
B41	Band 4.1 homologs	E(MP)	0(0)	19(20)	
BAG	BAG domains, present in regulator of Hsp70 proteins	E(MFP)	1(1)	2(2)	
BAH	Bromo adjacent homology domain	E(MFP)	5(5)	3(4)	
BBOX	B-Box-type zinc finger	E(MP)	0(0)	16(22)	1FRE

(continues)

TABLE I (*Continued*)

Domain	Definition	Phyletic distribution	Yeast proteins (domains)	Worm proteins (domains)	PDB
BCL	BCL (B-Cell lymphoma); contains BH1, BH2 regions	E(M)	0(0)	1(1)	1AF3
BH4	Bcl-2 homology region 4	E(M)	0(0)	1(1)	1AF3
BHL	Bacterial histone-like domain	E()AB	0(0)	0(0)	1HUE
BIR	Baculoviral inhibition of apoptosis protein repeat	E(MF)	1(1)	2(3)	
BPI1	BPI/LBP/CETP N-terminal domain	E(M)	0(0)	7(7)	1BP1
BPI2	BPI/LBP/CETP C-terminal domain	E(M)	0(0)	9(9)	1BP1
BRCT	Breast cancer carboxy-terminal domain	E(MFP)B	10(15)	28(40)	
BRLZ	Basic region leucine zipper	E(MFP)	15(15)	24(25)	1A02
BROMO	Bromo domain	E(MFP)	9(14)	15(23)	
BTB	Domain in Broad-Complex, Tramtrack and Bric a brac	E(MFP)	3(4)	130(138)	3KVT
btg1	Tob/Btg1 family	E(M)	0(0)	1(1)	
BTK	Bruton's tyrosine kinase Cys-rich motif	E(M)	0(0)	0(0)	1B55
B_lectin	Bulb-type mannose-specific lectin	E(P)B	0(0)	0(0)	1BWU
BowB	Bowman–Birk type proteinase inhibitor	E(P)	0(0)	0(0)	1SBW
C1	Protein kinase C conserved region 1 (C1) domains (cysteine-rich domains)	E(MFP)	1(2)	35(53)	1FAQ
C1Q	Complement component C1q domain.	E(M)	0(0)	0(0)	
C2	Protein kinase C conserved region 2 (CalB)	E(MFP)	11(22)	47(69)	1A25
C4	C-terminal tandem repeated domain in type 4 procollagens	E(M)	0(0)	2(4)	
CA	Cadherin repeats	E(M)B	0(0)	17(118)	1EDH
CAD	Domains present in proteins implicated in postmortem DNA fragmentation	E(M)	0(0)	0(0)	
CALCITONIN	Calcitonin	E(M)	0(0)	0(0)	1BKU
Calx_beta	Domain in Na-Ca exchangers and integrin-$\beta 4$	E(M)B	0(0)	3(6)	
CARD	Caspase recruitment domain	E(M)	0(0)	2(2)	3CRD

CASc	Caspase, interleukin-1β converting enzyme (ICE) homologs	E(M)	0(0)	3(3)	1BMQ
CBS	Domain in cystathionine β-synthase and other proteins.	E(MFP)AB	9(20)	9(20)	1B3O
CCP	Domain abundant in complement control proteins; SUSHI repeat; short complement-like repeat (SCR)	E(M)	0(0)	13(80)	1CKL
CH	Calponin homology domain	E(MFP)	4(7)	23(35)	1AA2
CheW	Two component signaling adaptor domain	AB	0(0)	0(0)	
CHROMO	Chromatin organization modifier domain	E(MFP)	4(5)	23(29)	1AP0
ChSh	Chromo shadow Domain	E(MFP)	0(0)	4(4)	
ChtBD	Chitin binding domain	E(MFP)	0(0)	3(7)	1HEV
CLECT	C-type lectin (CTL) or carbohydrate-recognition domain (CRD)	E(M)	0(0)	246(314)	1B6E
CLH	Clathrin heavy chain repeat homology	E(MFP)	3(9)	2(8)	1B89
CLa	Clusterin α chain	E(M)	0(0)	0(0)	
CLb	Clusterin β chain	E(M)	0(0)	0(0)	
CNH	Domain found in NIK1-like kinases, mouse citron, and yeast ROM1, ROM2	E(MF)	3(3)	5(5)	
cNMP	Cyclic nucleotide-monophosphate binding domain	E(MFP)AB	2(3)	17(23)	1APK
CNX	Connexin homologs	E(M)	0(0)	0(0)	
COLFI	Fibrillar collagens C-terminal domain	E(M)	0(0)	0(0)	
COLIPASE	Colipase	E(M)	0(0)	0(0)	1ETH
CRF	Corticotropin-releasing factor	E(M)	0(0)	0(0)	
CSF2	Granulocyte-macrophage colony-simulating factor (GM-CSF)	E(M)	0(0)	0(0)	1CSG
CSP	Cold shock protein domain	E(MP)B	0(0)	6(6)	1A62
CT	C-terminal cystine knot-like domain (CTCK)	E(M)	0(0)	1(1)	
Cu_FIST	Copper-Fist	E(F)	3(3)	0(0)	1CO4
CUB	Domain first found in C1r, Cls, uEGF, and bone morphogenetic protein	E(M)	0(0)	57(95)	1SFP
CULLIN	Cullin	E(MFP)	3(3)	6(6)	
CY	Cystatin-like domain	E(MP)	0(0)	2(2)	1A67

(*continues*)

TABLE I (*Continued*)

Domain	Definition	Phyletic distribution	Yeast proteins (domains)	Worm proteins (domains)	PDB
CYCLIN	Domain present in cyclins, TFIIB and retinoblastoma	E(MFP)A	18(27)	22(27)	1AIS
CYCc	Adenylyl/guanylyl cyclase, catalytic domain	E(MF)B	1(1)	36(39)	1AB8
CysPc	Calpain-like thiol protease family	E(MF)B	1(1)	14(14)	
DAGKa	Diacylglycerol kinase accessory domain (presumed)	E(MP)	0(0)	7(7)	
DAGKc	Diacylglycerol kinase catalytic domain (presumed)	E(MFP)B	2(2)	9(9)	
DAX	Domain present in Dishevelled and axin	E(M)	0(0)	3(3)	
DEATH	DEATH domain, found in proteins involved in cell death (apoptosis)	E(M)B	0(0)	9(9)	1DDF
DED	Death effector domain	E(M)	0(0)	0(0)	1A1W
DEFSN	Defensin/corticostatin family	E(M)	0(0)	0(0)	1BNB
DEP	Domain found in Dishevelled, Egl-10, and Pleckstrin	E(MF)	5(5)	10(10)	
DISIN	Homologs of snake disintegrins	E(MF)	0(0)	5(5)	1FVL
DM	Doublesex DNA-binding motif	E(M)	0(0)	10(12)	
DnaJ	DnaJ molecular chaperone homology domain	E(MFP)AB	21(21)	36(36)	1BQ0
DSL	Delta serrate ligand	E(M)	0(0)	9(9)	
DSRM	Double-stranded RNA binding motif	E(MFP)B	2(2)	12(19)	1STU
DUF1	Domain of unknown function with GGDEF motif	B	0(0)	0(0)	
DUF2	Domain of unknown function 2	B	0(0)	0(0)	
DYNc	Dynamin, GTPase	E(MFP)	3(3)	4(4)	
EFh	EF-hand, calcium binding motif	E(MFP)B	9(27)	53(163)	1B7T
EGF	Epidermal growth factor (EGF)-like domain	E(MFP)	0(0)	62(158)	1A3P
EGF_CA	Calcium-binding EGF-like domain	E(MP)	0(0)	29(65)	1F7E
EGF_Lam	Laminin-type EGF-like domain	E(M)	0(0)	13(96)	1KLO
EGF_like	EGF domain, unclassified subfamily	E(MFP)	1(1)	57(289)	1DAN
EH	Eps15 homology domain	E(MF)	5(9)	2(3)	
END	Endothelin	E(M)	0(0)	0(0)	1EDN
ENTH	Epsin N-terminal homology (ENTH) domain	E(MFP)	8(8)	5(5)	

EPEND	Ependymins	E(M)	0(0)	0(0)	
ETS	Erythroblast transformation specific domain	E(M)	0(0)	10(11)	1AWC
FA58C	Coagulation factor 5/8 C-terminal domain, discoidin domain	E(MF)B	0(0)	3(3)	1EUT
FBG	Fibrinogen-related domains (FReDs)	E(M)	0(0)	6(6)	1FZE
FBOX	A receptor for ubiquitination targets	E(MFP)	9(9)	156(175)	
fCBD	Fungal-type cellulose-binding domain	E(F)	0(0)	0(0)	1AZ6
FCH	Fes/CIP4 homology domain	E(MF)	4(4)	6(6)	
FGF	Acidic and basic fibroblast growth factor family	E(M)	0(0)	1(1)	1QCT
FH	Forkhead domain	E(MF)	4(4)	18(18)	2HFH
FHA	Forkhead associated domain	E(MFP)B	13(14)	12(12)	
FIMAC	Factor I membrane attack complex	E(M)	0(0)	0(0)	
FN1	Fibronectin type 1 domain	E(M)	0(0)	0(0)	1FBR
FN2	Fibronectin type 2 domain	E(M)	0(0)	0(0)	1PDC
FN3	Fibronectin type 3 domain	E(MFP)AB	2(2)	49(212)	1QR4
FOLN	Follistatin N-terminal domain-like	E(M)	0(0)	4(18)	1BMO
FRI	Frizzled	E(M)	0(0)	4(4)	
FU	Furin-like repeats	E(M)	0(0)	3(12)	
FYVE	Protein present in Fab1, YOTB, Vac1, and EEA1	E(MFP)	5(6)	15(15)	1VFY
G-alpha	G protein α subunit	E(MFP)	2(2)	20(20)	1CIP
GAF	Domain present in phytochromes and cGMP-specific phosphodiesterases	E(MFP)AB	1(1)	2(2)	
GAL4	GAL4-like Zn(II)2Cys6 (or C6 zinc) binuclear cluster DNA-binding domain	E(F)	54(54)	0(0)	1HWT
Galanin	Galanin	E(M)	0(0)	0(0)	
GAS2	Growth-Arrest-Specific protein 2 Domain	E(M)	0(0)	3(3)	
GASTRIN	Gastrin/cholecystokinin/caerulein family	E(M)	0(0)	0(0)	
GED	Dynamin GTPase effector domain	E(MFP)	2(2)	3(3)	
GEL	Gelsolin homology domain	E(MP)	0(0)	3(11)	1SVQ
GGL	G protein γ subunit-like motifs	E(MFP)	1(1)	4(4)	1A0R
GHA	Glycoprotein hormone α chain homologs	E(M)	0(0)	0(0)	1HCN

(continues)

TABLE I (*Continued*)

Domain	Definition	Phyletic distribution	Yeast proteins (domains)	Worm proteins (domains)	PDB
GHB	Glycoprotein hormone β chain homologs	E(M)	0(0)	1(1)	1HCN
GLA	Domain containing Gla (γ-carboxyglutamate) residues	E(M)	0(0)	0(0)	1CFH
GLECT	Galectin	E(MF)	0(0)	25(33)	1A3K
GLUCA	Glucagon	E(MP)	0(0)	0(0)	1BH0
GPS	G-protein-coupled receptor proteolytic site domain	E(M)	0(0)	7(7)	
GRAN	Granulin	E(MP)	0(0)	1(3)	
GuKc	Guanylate kinase homologs	E(MFP)B	1(1)	8(8)	1GKY
H2A	Histone 2A	E(MFP)	3(3)	15(15)	1AOI
H2B	Histone H2B	E(MFP)	2(2)	12(12)	1AOI
H3	Histone H3	E(MFP)	3(3)	22(22)	1AOI
H4	Histone H4	E(MFP)	2(2)	13(13)	1AOI
HALZ	Homeobox associated leucine zipper	E(P)	0(0)	0(0)	
HAMP	HAMP (Histidine kinases, Adenylyl cyclases, Methyl binding proteins, Phosphatases) domain	E(MF)AB	0(0)	0(0)	
HAT	HAT (Half-A-TPR) repeats	E(MFP)	8(53)	7(65)	
HATPase_c	Histidine kinase-like ATPases	E(MFP)AB	8(8)	9(9)	1A4H
HECTc	Domain homologous to E6-AP Carboxyl Terminus	E(MFP)	5(5)	7(7)	
HhH1	Helix–hairpin–helix DNA-binding motif class 1	E(MFP)AB	2(2)	2(3)	1BDX
HhH2	Helix–hairpin–helix class 2 (Pol1 family) motifs	E(MFP)AB	0(0)	4(4)	1BGX
HintC	Hint (Hedgehog/Intein) domain C-terminal region	E(MF)AB	1(1)	10(10)	1AM2
HintN	Hint (Hedgehog/Intein) domain N-terminal region	E(MFP)AB	2(2)	9(9)	1AM2
HisKA	His Kinase A (phosphoacceptor) domain	E(MFP)AB	1(1)	1(1)	
HLH	Helix loop helix domain	E(MFP)	7(7)	35(40)	1A0A
HLT	Histone-like transcription factor	E(MFP)	2(2)	2(2)	
HMG	High mobility group	E(MFP)	7(9)	15(17)	1AAB

HormR	Domain present in hormone receptors	E(M)	0(0)	5(5)	
HOX	Homeodomain	E(MFP)	10(10)	92(99)	1AHD
HPT	Histidine phosphotransfer domain	E(FP)AB	1(1)	0(0)	1BDJ
HR1	Protein kinase C-related kinase homology region 1 homologs	E(MF)	1(2)	1(1)	
HRDC	Helicase and RNase D C-terminal	E(MFP)B	2(2)	3(3)	
HSF	Heat shock factor	E(MFP)B	5(5)	0(0)	1HKS
HTH_ARAC	helix_turn_helix, arabinose operon control protein	E(M)B	0(0)	0(0)	1BL0
HTH_ARSR	helix_turn_helix, Arsenical Resistance Operon Repressor	AB	0(0)	0(0)	1SMT
HTH_ASNC	helix_turn_helix ASNC type	AB	0(0)	0(0)	
HTH_CRP	helix_turn_helix, cAMP Regulatory protein	E()AB	0(0)	0(0)	1BER
HTH_DEOR	helix_turn_helix, Deoxyribose operon repressor	B	0(0)	0(0)	
HTH1_GNTR	helix_turn_helix gluconate operon transcriptional repressor	E(M)B	0(0)	0(0)	
HTH_ICLR	helix_turn_helix isocitrate lyase regulation	B	0(0)	0(0)	
HTH_LACI	helix_turn_helix lactose operon repressor	B	0(0)	0(0)	1QP0
HTH_LUXR	helix_turn_helix, Lux regulon	E()B	0(0)	0(0)	1A04
HTH_MARR	helix_turn_helix multiple antibiotic resistance protein	AB	0(0)	0(0)	
HTH_MERR	helix_turn_helix, mercury resistance	E(M)AB	0(0)	0(0)	
HX	Hemopexin-like repeats	E(MP)B	0(0)	4(9)	1FBL
HYDRO	Hydrophobins	E(F)	0(0)	0(0)	
IB	Insulin growth factor-binding protein homologs	E(M)	0(0)	0(0)	1BOE
IFabd	Interferon α, β, and δ	E(M)	0(0)	0(0)	1B5L
IG	Immunoglobulin	E(M)B	0(0)	17(21)	1A64
IG_like	Immunoglobulin-like	E(M)B	0(0)	50(220)	12E8
IGc1	Immunoglobulin C-type	E(M)	0(0)	0(0)	1A1M
IGc2	Immunoglobulin C-2-type	E(M)	0(0)	53(231)	1ALS
IGv	Immunoglobulin V-type	E(M)	0(0)	0(0)	12E8
IL1	Interleukin-1 homologs	E(M)	0(0)	0(0)	1HIB
IL10	Interleukin-10 family	E(M)	0(0)	0(0)	1ILK
IL2	Interleukin-2 family	E(M)	0(0)	0(0)	1ILM

(*continues*)

TABLE I (*Continued*)

Domain	Definition	Phyletic distribution	Yeast proteins (domains)	Worm proteins (domains)	PDB
IL4_13	Interleukins 4 and 13	E(M)	0(0)	0(0)	1BCN
IL6	Interleukin-6 homologs	E(M)	0(0)	0(0)	1ALU
IL7	Interleukin-7 and interleukin-9 family.	E(M)	0(0)	0(0)	1IL7
ILWEQ	I/LWEQ domain	E(MF)	1(1)	3(3)	
INB	Integrin β subunits (N-terminal portion of extracellular region)	E(M)	0(0)	2(2)	
Int_alpha	Integrin α (β-propeller repeats)	E(MF)B	1(1)	2(9)	1A8X
IPPc	Inositol polyphosphate phosphatase, catalytic domain homologs	E(MFP)	4(4)	5(5)	
IQ	Short calmodulin-binding motif containing conserved Ile and Gln residues	E(MFP)B	4(11)	16(32)	1B7T
IRF	Interferon regulatory factor	E(M)	0(0)	0(0)	1IF1
ITAM	Immunoreceptor tyrosine-based activation motif	E(M)	0(0)	0(0)	
IlGF	Insulin/insulin-like growth factor/relaxin family	E(M)	0(0)	10(0)	1BQT
JAB_MPN	JAB/MPN domain	E(MFP)AB	4(4)	7(7)	
KAZAL	Kazal type serine protease inhibitors	E(M)	0(0)	3(26)	1AN1
KH	K homology RNA-binding domain	E(MFP)AB	9(19)	29(68)	1VIG
KISc	Kinesin motor, catalytic domain, ATPase	E(MFP)	6(6)	20(20)	2NCD
KR	Kringle domain	E(M)	0(0)	3(3)	1KIV
KRAB	Krueppel associated box	E(M)	0(0)	0(0)	
KU	BPTI/Kunitz family of serine protease inhibitors	E(M)	0(0)	42(141)	1AAL
LamB	Laminin B domain	E(M)	0(0)	7(12)	
LamG	Laminin G domain	E(M)B	0(0)	16(36)	1SLI
LamNT	Laminin N-terminal domain (domain VI)	E(M)	0(0)	5(5)	
LDLa	Low-density lipoprotein receptor domain class A	E(MFP)	1(1)	34(147)	1AJJ
LGN	LGN motif, putative GEFs specific for G-α GTPases	E(M)	0(0)	3(6)	
LH2	Lipoxygenase homology 2 (β barrel) domain	E(MP)	0(0)	2(2)	1BU8

LIF_OSM	Leukemia inhibitory factor	E(M)	0(0)	0(0)	1A7M
LIM	Zinc-binding domain present in Lin-11, Isl-1, Mec-3.	E(MFP)	4(9)	32(73)	1B8T
LINK	Link (Hyaluronan-binding)	E(M)	0(0)	1(1)	1TSG
LMWPc	Low molecular weight phosphatase family	E(MF)AB	1(1)	0(0)	1BVH
LRR	Leucine-rich repeats	E(MFP)B	11(63)	73(403)	1A4Y
LRRCT	Leucine rich repeat C-terminal domain	E(M)	0(0)	7(9)	
LRRNT	Leucine rich repeat N-terminal domain	E(M)	0(0)	3(6)	
LU	Ly-6 antigen/uPA receptor-like domain	E(M)	0(0)	3(3)	1CDQ
LY	Low-density lipoprotein-receptor YWTD domain	E(M)	0(0)	7(72)	1LPX
LysM	Lysin motif	E(MFP)B	0(0)	5(21)	
LYZ1	α-Lactalbumin/lysozyme C	E(M)	0(0)	0(0)	1A4V
LYZ2	Lysozyme subfamily 2	B	0(0)	0(0)	
MA	Methyl-accepting chemotaxis-like domains (chemotaxis sensory transducer)	E(M)AB	0(0)	0(0)	
MACPF	Conserved domain in membrane attack complex proteins and perforin	E(MP)B	0(0)	0(0)	
MAM	Domain in meprin, A5, receptor protein tyrosine phosphatase mu (and others)	E(M)	0(0)	1(1)	
MATH	Meprin and TRAF homology	E(MFP)	1(1)	87(155)	1CA4
MBD	Methyl-CpG binding domain	E(MP)	0(0)	2(2)	
MCM	Minichromosome maintenance proteins	E(MFP)A	6(6)	6(6)	
MeTrc	Methyltransferase, chemotaxis proteins	E(F)AB	0(0)	0(0)	1AF7
MyTH4	Domain in Myosin and Kinesin Tails	E(MP)	0(0)	2(4)	
MYSc	Myosin, Large ATPases	E(MFP)	5(5)	17(17)	1B7T
NAT_PEP	Natriuretic peptide	E(M)	0(0)	0(0)	1ANP
NEBU	Nebulin repeats	E(M)	0(0)	1(2)	
NGF	Nerve growth factor (NGF or β-NGF)	E(M)	0(0)	0(0)	1NT3
NH	Neurohypophysial hormones	E(M)	0(0)	0(0)	1NPO
NL	Domain found in Notch and Lin-12	E(M)	0(0)	2(6)	
NMU	Neuromedin U	E(M)	0(0)	0(0)	
OLF	Olfactomedin-like domains	E(M)	0(0)	2(2)	

(continues)

TABLE I (*Continued*)

Domain	Definition	Phyletic distribution	Yeast proteins (domains)	Worm proteins (domains)	PDB
OPR	Octicosapeptide repeat	E(MFP)	1(1)	3(3)	
OSTEO	Osteopontin	E(M)	0(0)	0(0)	
P	P or trefoil or TFF domain	E(M)	0(0)	2(2)	1PCP
PA2	Phospholipase A2	E(MP)	0(0)	2(2)	1CL5
PAC	Motif C-terminal to PAS motifs (likely to contribute to PAS structural domain)	E(MFP)AB	0(0)	5(5)	2ARN
PAH	Pancreatic hormones/neuropeptide F/peptide YY family	E(M)	0(0)	0(0)	1BBA
PAS	PAS domain	E(MFP)AB	3(4)	9(14)	2ARN
PAX	Paired box domain	E(M)	0(0)	9(9)	1PDN
PBD	P21-Rho-binding domain	E(MFP)	5(5)	0(0)	
PBPb	Bacterial periplasmic substrate-binding proteins	E(M)AB	0(0)	0(0)	1GGG
PBPe	Eukaryotic homologs of bacterial periplasmic substrate binding proteins	E(MP)	0(0)	8(8)	1GR2
PDEc	Cyclic nucleotide phosphodiesterase catalytic domain	E(MF)	1(1)	6(6)	
PDGF	Platelet-derived and vascular endothelial growth factors (PDGF, VEGF) family	E(M)	0(0)	1(1)	1BJ1
PDZ	Domain present in PSD-95, Dlg, and ZO-1/2	E(MFP)AB	2(3)	64(87)	1B8Q
PH	Pleckstrin homology domain	E(MFP)	28(31)	76(83)	1B55
PHB	Prohibitin homologs	E(MFP)AB	2(2)	12(12)	
PHD	PHD zinc finger	E(MFP)	15(19)	35(51)	
PI3K_C2	Phosphoinositide 3-kinase, region postulated to contain C2 domain	E(MFP)	1(1)	2(2)	1QMM
PI3K_p85B	PI3-kinase family, p85-binding domain	E(M)	0(0)	0(0)	
PI3K_rbd	PI3-kinase family, Ras-binding domain	E(M)	0(0)	1(1)	1QMM
PI3Ka	Phosphoinositide 3-kinase family, accessory domain (PIK domain)	E(MFP)	2(2)	3(3)	1QMM
PI3Kc	Phosphoinositide 3-kinase, catalytic domain	E(MFP)	8(8)	9(9)	1QMM

PINT	Motif in proteasome subunits, Int-6, Nip-1 and TRIP-15	E(MFP)	8(8)	12(12)	
PIPKc	Phosphatidylinositol phosphate kinases	E(MFP)	2(2)	3(3)	1BO1
PKD	Repeats in polycystic kidney disease 1 (PKD1) and other proteins	E(M)AB	0(0)	0(0)	1B4R
PLAc	Cytoplasmic phospholipase A_2, catalytic subunit	E(MF)	4(4)	0(0)	
PLCXc	Phospholipase C, catalytic domain (part); domain X	E(MFP)B	1(1)	6(6)	1AOD
PLCYc	Phospholipase C, catalytic domain (part); domain Y	E(MFP)	1(1)	6(6)	1DJG
PLDc	Phospholipase D. Active site motifs.	E(MFP)B	1(2)	5(9)	
PLEC	Plectin repeat	E(M)	0(0)	1(16)	
PLP	Myelin proteolipid protein (PLP or lipophilin)	E(M)	0(0)	0(0)	
POU	Found in Pit-Oct-Unc transcription factors	E(M)	0(0)	4(4)	1AU7
PP2Ac	Protein phosphatase 2A homologs, catalytic domain	E(MFP)AB	12(12)	47(47)	1AUL
PP2C_SIG	Sigma factor PP2C-like phosphatases	E(MFP)B	6(6)	7(7)	1A6Q
PP2Cc	Serine/threonine phosphatases, family 2C, catalytic domain	E(MFP)B	8(8)	11(11)	1A6Q
PROF	Profilin	E(MFP)	1(1)	3(3)	1A0K
PRP	Major prion protein	E(M)	0(0)	0(0)	1AG2
PSI	Domain found in Plexins, Semaphorins and Integrins	E(M)	0(0)	6(12)	
PTB	Phosphotyrosine-binding domain, phosphotyrosine-interaction (PI) domain	E(M)	0(0)	11(11)	1AQC
PTBI	Phosphotyrosine-binding domain (IRSI-like)	E(M)	0(0)	1(1)	1IRS
PTH	Parathyroid hormone	E(M)	0(0)	0(0)	1BZG
PTI	Plant trypsin inhibitors	E(P)	0(0)	0(0)	1MCU
PTN	Pleiotrophin/midkine family	E(M)	0(0)	0(0)	1MKC
DSPc	Dual specificity phosphatase, catalytic domain	E(MFP)B	5(5)	10(10)	1VHR
PTPc	Protein tyrosine phosphatase, catalytic domain	E(MFP)B	3(3)	85(88)	1A5Y
PTPc_DSPc	Protein tyrosine phosphatase, catalytic domain, undefined specificity	E(MFP)AB	3(3)	17(17)	1BZC
PTX	Pentraxin/C-reactive protein/pentaxin family	E(M)	0(0)	0(0)	1CRV
PUA	Putative RNA-binding domain in PseudoUridine synthase and Archaeosine transglycosylase	E(MFP)AB	5(5)	3(3)	

(*continues*)

TABLE I (*Continued*)

Domain	Definition	Phyletic distribution	Yeast proteins (domains)	Worm proteins (domains)	PDB
Pumilio	Pumilio-like repeats	E(MFP)	7(50)	12(80)	
PWI	PWI domain in splicing factors	E(MFP)	1(1)	3(3)	
PWWP	domain with conserved PWWP motif	E(MFP)	2(2)	1(1)	
PX	PhoX homologous domain, present in p47phox and p40phox	E(MFP)	14(14)	10(10)	
PXA	Domain associated with PX domains	E(MF)	1(1)	1(1)	
R3H	Putative single-stranded nucleic acids-binding domain	E(MFP)B	2(2)	3(3)	
RA	Ras association (RalGDS/AF-6) domain	E(MF)	2(2)	9(12)	1LFD
RAB	Rab subfamily of small GTPases	E(MFP)	9(9)	24(24)	3RAB
RAN	Ran (Ras-related nuclear proteins)/TC4 subfamily of small GTPases	E(MFP)	2(2)	1(1)	1IBR
RanBD	Ran-binding domain	E(MFP)	3(3)	2(3)	1EVH
RAS	Ras subfamily of RAS small GTPases	E(MF)	3(3)	8(8)	1CLU
RasGAP	GTPase-activator protein for Ras-like GTPases	E(MF)	3(3)	2(2)	1WER
RasGEF	Guanine nucleotide exchange factor for Ras-like small GTPases	E(MF)	5(5)	8(8)	1BKD
RasGEFN	Guanine nucleotide exchange factor for Ras-like GTPases; N-terminal motif	E(MF)	5(5)	7(7)	1BKD
REC	Che Y-homologous receiver domain	E(MFP)AB	4(4)	0(0)	1BDJ
RGS	Regulator of G protein signaling domain	E(MF)	3(3)	14(15)	1AGR
RHO	Rho (Ras homology) subfamily of Ras-like small GTPases	E(MFP)	6(6)	7(7)	1CF4
RhoGAP	GTPase-activator protein for Rho-like GTPases	E(MFP)B	11(11)	21(21)	1AM4
RhoGEF	Guanine nucleotide exchange factor for Rho/Rac/Cdc42-like GTPases	E(MF)	6(6)	19(20)	1DBH
RIIa	RIIalpha, Regulatory subunit portion of type II PKA R-subunit	E(MF)	1(1)	2(2)	1APK
RING	Ring finger	E(MFP)	38(41)	128(140)	1BOR

RIO	RIO-like kinase	E(MFP)A	2(2)	2(2)	
RNase_Pc	Pancreatic ribonuclease	E(M)	0(0)	0(0)	1B6V
RRM	RNA recognition motif	E(MFP)	60(101)	109(165)	2UP1
S1	Ribosomal protein S1-like RNA-binding domain	E(MFP)AB	7(18)	6(14)	1AH9
S4	S4 RNA-binding domain	E(MFP)AB	8(8)	2(2)	2TS1
SAA	Serum amyloid A proteins	E(M)	0(0)	0(0)	
SAM	Sterile α motif	E(MFP)	4(4)	17(24)	1B0X
SAM_PNT	SAM/Pointed domain	E(M)	0(0)	2(2)	1BQV
SAND	SAND domain	E(M)	0(0)	4(4)	
SANT	SW13, ADA2, N-CoR and TFIIIB" DNA-binding domains	E(MFP)B	18(31)	18(31)	1A5J
SAPA	Saposin/surfactant protein-B A-type DOMAIN	E(M)	0(0)	0(0)	
SAPB	Saposins-like type B	E(MP)	0(0)	20(25)	1NKL
SAR	Sar1p-like members of the Ras-family of small GTPases	E(MFP)	1(1)	1(1)	
SCP	SCP/Tpx-1/Ag5/PR-1/Sc7 family of extracellular domains	E(MFP)	3(3)	35(36)	1CFE
SCY	Intercrine alpha family (small cytokine C-X-C) (chemokine CXC)	E(M)	0(0)	0(0)	1B2T
SEA	Domain found in sea urchin sperm protein, enterokinase, agrin	E(M)	0(0)	4(6)	
Sec7	Sec7 domain	E(MFP)B	5(5)	5(5)	1BC9
SERPIN	SERine Proteinase Inhibitors	E(MP)	0(0)	10(10)	1A7C
SET	SET (Su(var)3-9, Enhancer-of-zeste, Trithorax) domain	E(MFP)B	7(7)	27(28)	
SF_P	Domain in pulmonary surfactant proteins	E(M)	0(0)	0(0)	
SH2	Src homology 2 domains	E(MFP)	1(1)	63(65)	1QCF
SH3	Src homology 3 domains	E(MFP)B	25(29)	61(74)	1AZE
SH3b	Bacterial SH3 domain homologs	B	0(0)	0(0)	
ShKT	ShK toxin domain	E(MP)	0(0)	100(259)	1BEI
SNc	Staphylococcal nuclease homologs	E(MFP)AB	1(1)	1(4)	1A2T
small_GTPase	Small GTPase of the Ras superfamily; ill-defined subfamily	E(MFP)AB	11(13)	34(34)	1SCG
SO	Somatomedin B -like domains	E(M)	0(0)	1(1)	
SOCS	Suppressors of cytokine signaling	E(M)	0(0)	2(2)	
SPEC	Spectrin repeats	E(MP)	0(0)	14(145)	1AJ3

(*continues*)

TABLE I (*Continued*)

Domain	Definition	Phyletic distribution	Yeast proteins (domains)	Worm proteins (domains)	PDB
SPRY	Domain in SPla and the RYanodine Receptor	E(MFP)	3(3)	10(12)	
SR	Scavenger receptor Cys-rich	E(MP)	0(0)	1(2)	1BY2
START	In StAR and phosphatidylcholine transfer protein	E(MP)	0(0)	6(6)	
STE	STE-like transcription factors	E(F)	1(1)	0(0)	
SWIB	SWI complex, BAF60b domains	E(MFP)B	2(2)	3(3)	
STYKc	Protein kinase; unclassified specificity	E(MFP)AB	5(5)	92(95)	
S_TKc	Serine/threonine protein kinases, cytalytic domain	E(MFP)B	112(112)	256(260)	1A9U
S_TK_X	Extension to Ser/Thr-type protein kinases	E(MFP)	12(12)	27(28)	1BX6
TarH	Homologs of the ligand binding domain of Tar	B	0(0)	0(0)	1LIH
TBC	Domain in Tre-2, BUB2p, and Cdc16p Probable Rab-GAPs	E(MFP)	10(10)	20(20)	
TBOX	Domain first found in the mice T locus (brachyury) protein	E(M)	0(0)	21(21)	1XBR
TEA	TEA domain	E(MF)	1(1)	1(1)	
TGFB	Transforming growth factor-β(TGF-β) family	E(M)	0(0)	4(4)	1AGQ
THN	Thaumatin family	E(MP)	0(0)	6(6)	1AUN
THY	Thymosin β actin-binding motif	E(M)	0(0)	1(3)	
TIMP	Tissue inhibitor of metalloproteinase family.	E(M)	0(0)	1(1)	1BR9
TIR	Toll-interleukin 1-resistance	E(MP)B	0(0)	2(2)	
TK	Tachykinin family	E(M)	0(0)	0(0)	
TNF	Tumor necrosis factor family	E(M)	0(0)	0(0)	5TSW
TNFR	Tumor necrosis factor receptor/nerve growth factor receptor repeats	E(MP)	0(0)	1(1)	1CDF
TPR	Tetratricopeptide repeats	E(MFP)AB	18(92)	30(139)	1A17
TR_FER	Transferrin	E(MP)	0(0)	0(0)	1A8E
TR_THY	Transthyretin	E(MF)B	0(0)	1(1)	5TTR
Tryp_SPc	Trypsin-like serine protease	E(MFP)B	0(0)	13(13)	1A5H
TSP1	Thrombospondin type 1 repeats	E(M)	0(0)	27(119)	

TSPN	Thrombospondin N-terminal-like domains	E(M)	0(0)	0(0)	
TSPc	Tail specific protease	E(MP)AB	0(0)	0(0)	
TUDOR	Tudor domain	E(MFP)	0(0)	8(11)	
TY	Thyroglobulin type I repeats	E(M)	0(0)	6(12)	
TyrKc	Tyrosine kinase, catalytic domain	E(MP)	0(0)	73(73)	1QCF
t_SNARE	Helical region found in SNAREs	E(MFP)	14(15)	13(15)	1SFC
UBA	Ubiquitin associated domain	E(MFP)	8(10)	9(9)	1UBA
UBCc	Ubiquitin-conjugating enzyme E2, catalytic domain homologs	E(MFP)	15(15)	22(22)	1QCQ
UBQ	Ubiquitin homologs	E(MFP)	10(14)	22(33)	1UD7
UBX	Domain present in ubiquitin-regulatory proteins	E(MFP)	7(7)	2(2)	
UTG	Uteroglobin	E(M)	0(0)	0(0)	1CCD
VHP	Villin headpiece domain	E(MP)	0(0)	3(3)	1VII
VHS	Domain present in VPS-27, Hrs and STAM	E(MFP)	4(4)	4(4)	
VPS9	Domain present in VPS9	E(MF)	2(2)	3(3)	
VWA	von Willebrand factor (vWF) type A domain	E(MFP)AB	4(4)	54(60)	1A8X
VWC	von Willebrand factor (vWF) type C domain	E(M)	0(0)	3(8)	
VWD	von Willebrand factor (vWF) type D domain	E(M)	0(0)	9(9)	
WAP	Four-disulfide core domains	E(M)	0(0)	5(13)	1CJH
WD40	WD40 repeats	E(MFP)AB	101(514)	129(677)	1A0R
WH1	WASP homology region 1	E(MF)	1(1)	2(3)	1EVH
WH2	Wiskott-Aldrich syndrome homology region 2	E(MF)	2(3)	4(6)	
WNT1	found in Wnt-1	E(M)	0(0)	5(5)	
WR1	Worm-specific repeat type 1	E(M)	0(0)	41(222)	
WSC	Domain present in yeast cell wall integrity and stress response component proteins	E(MF)	4(4)	0(0)	
WW	Domain with 2 conserved Trp (W) residues	E(MFP)	6(9)	15(24)	1PIN
XTALbg	β/γ crystallins	E(M)B	0(0)	0(0)	1A45
ZnF_A20	A20-like zinc fingers	E(MP)	0(0)	2(2)	
ZnF_AN1	AN1-like zinc finger	E(MFP)	2(2)	3(4)	

(continues)

TABLE I (*Continued*)

Domain	Definition	Phyletic distribution	Yeast proteins (domains)	Worm proteins (domains)	PDB
ZnF_C2H2	Zinc finger	E(MFP)AB	50(113)	198(643)	1BHI
ZnF_C2HC	Zinc finger	E(MFP)B	13(28)	43(65)	1NC8
ZnF_C3H1	Zinc finger	E(MFP)	6(14)	28(58)	
ZnF_C4	C4 zinc finger in nuclear hormone receptors	E(M)	0(0)	248(253)	1A6Y
ZnF_CHCC	Zinc finger	B	0(0)	0(0)	
ZnF_GATA	Zinc finger binding to DNA consensus sequence [AT]GATA[AG]	E(MFP)	10(10)	12(14)	1GAT
ZnF_UBP	Ubiquitin carboxyl-terminal hydrolase-like zinc finger	E(MFP)	4(4)	4(4)	
ZnF_UBR1	Putative zinc finger in N-recognin, a recognition component of the N-end rule pathway	E(MFP)	2(2)	6(6)	
ZnF_ZZ	Zinc-binding domain, present in Dystrophin, CREB-binding protein	E(MFP)	2(2)	11(12)	
ZnMc	Zinc-dependent metalloprotease	E(MP)B	0(0)	44(44)	1A85
ZP	Zona pellucida domain	E(M)	0(0)	34(34)	
ZU5	Domain present in ZO-1 and Unc5-like netrin receptors	E(M)	0(0)	6(6)	

[a] The phyletic distributions of families (E, eukaryota; M, metazoa; F, fungi; P, Viridiplantae (plants); B, bacteria, A., archaea) and the numbers of proteins (domains) detected in the *S. cerevisiae* ("yeast") and *C. elegans* ("worm") genomes are shown. The rightmost column contains a representative PDB code for determined tertiary structures of the domain family, if known.

function. Regions predicted to form coiled coils (Lupas, 1997) and yielding apparently significant *E* values are treated with extreme caution, as sequence similarities between such structures are unlikely to be biologically meaningful. Sequence database searches employ both a nonredundant protein sequence database (nrdb) (ftp://ncbi.nlm.nih.gov/blast/db/nr) and a nrdb with no sequence pairs with greater than 90% sequence identity (Holm and Sander, 1998; ftp://ftp.ebi.ac.uk/pub/databases/nrdb90).

Homologs are identified in an iterative search protocol. The initial multiple alignment may be derived from structure-based alignments of divergent homologs (Holm and Sander, 1996) where available, or from Clustal derived (Thompson *et al.*, 1994) alignments of homologs identified by PSI-BLAST analysis. However, multiple alignments are always manually edited to ensure optimization (cf. Bork and Gibson, 1996). This includes the removal of unnecessary insertion/deletion positions and optimal conservation of hydrophobic or polar residues within known or predicted secondary structures. Hypothetical proteins predicted from genomic sequence that appear to be misassembled are deleted from these alignments. Domain limits are assessed from known structures, *bona fide* protein N and C termini or from the known limits of adjacent domains. Alignments are rigorously inspected for nonconservation within otherwise well-conserved blocks that indicate the inclusion of false-positive sequences or true-positive sequences containing sequence errors. One of all pairs of sequences with greater than 67% pairwise sequence identity is purged from the alignment. This reduces the size of the alignment and assists in ensuring that similar sequences are not overrepresented.

Typically, a HMM, prepared from this alignment, is then compared with current sequence databases. Simultaneously, each sequence from the alignment is used as a query in PSI-BLAST searches. All sequences aligned with significant scores against the HMM or PSI-BLAST profile are collected and realigned, as described previously, to proceed with the subsequent iteration. This procedure is followed until no new putative homologs are detected. New alignments are constructed, not via the pairwise method of CLUSTAL, but using the sequence-versus-profile/HMM method of the hmmalign algorithm of HMMER (Eddy, S., unpublished). Thus all the resulting sequences are related, either directly or indirectly, by significant *E* values in database searches. The SMART database stores the final multiple alignment, the highest *E* value of identified true positives (E_p), the lowest *E* value of predicted true negatives (E_n), and the size of the database searched. The latter is used to scale *E* value thresholds to ensure that identification of homologs is independent of database size. SMART will predict a domain homolog within any sequence that, when aligned with the relevant HMM using HMMER2 (S. Eddy, unpublished), yields an E value lower than E_p or if the E value lies between E_p and E_n and is less than 1.0.

The construction of the cold shock protein (**CSP**)[1] domain alignment for SMART is presented as an illustration of this process. The **CSP** domain family is represented throughout the bacteria and eukarya and appears to possess RNA chaperone functions (Graumann and Marahiel, 1998). An alignment was constructed (Thompson *et al.*, 1994) of all **CSP** homologs detectable by PSI-BLAST ($E < 0.01$) as significantly similar to the sequence of the known structure (Schindelin *et al.*, 1994) of *Escherichia coli* cold shock protein. A HMM was constructed from this alignment using HMMER2's **hmmbuild** algorithm and default parameters. Using this HMM to search nrdb90 (Holm and Sander, 1998; ftp://ftp.ebi.ac.uk/pub/databases/nrdb90) revealed additional known homologs with E values less than 0.1. In a subsequent iteration, *Thiobacillus ferrooxidans* VacB, a RNase II, was identified with $E = 8.8 \times 10^{-2}$ as a putative **CSP** domain homolog. This relationship was not revealed by a recent survey of ribonucleases (Mian, 1997). Two further iterations revealed a domain similar to **CSP** domains in Rho transcription termination factors. Although not significant according to the criteria described previously (lowest $E = 0.6$), these sequences were considered **CSP** domain homologs, as the known structures of *E. coli* Rho demonstrate substantial structural and functional similarities to **CSP** domains (Allison *et al.*, 1998; Briercheck *et al.*, 1998). Consequently, they were assigned as **CSP** domain homologs within a multiple alignment, whose corresponding HMM was unable to detect further examples of this family. S1-like RNA-binding domains (**S1**) were detected in HMMER2 database searches (lowest $E = 1.8$) and as distantly similar sequences in PSI-BLAST searches (data not shown). These domains also possess an OB fold (Bycroft *et al.*, 1997) and function common to **CSP** domains and Rho domains, and hence are likely distant homologs of this family. However, for the purposes of the SMART database, the **S1** family is being maintained as a separate family.

2. *Sequence Repeat Characterization*

Sequence repeats associate to form one of three broad classes of structure: a linear rod containing repeats arranged in an end-to-end

[1] To facilitate cross referencing between the names of domain families used in this article and structural, functional, and evolution information available from the literature, the domain names used by the WWW-based resource SMART (http://smart.embl-heidelberg) are shown in bold and in a proportional font.

manner (for example, spectrin repeats), a superhelix (for example, tetratrico peptide repeats [**TPRs**]), or a "closed" structure with interactions between the N- and C-terminal repeats (for example, **WD40** repeats in a β-propeller arrangement) (Fig. 3, see Color insert). The latter "closed" structures are compact and usually possess a hydrophobic core, and so each set of these repeats may be termed a domain. However, since recognition of repeats poses a different challenge from the recognition of domains, their detection requires a protocol that differs from that of domains.

Sequence repeats are observed within many protein families and many diverse organisms. At least 3% of eukaryotic proteins contain recognizable repeats (Andrade *et al.*, 1999b). Detection of sequence repeats is often more complicated than that of domains. They are extremely divergent with the result that it is often difficult to distinguish related repeats from phylogenetically unrelated regions. This can be countered by exploiting the characteristic that repeats co-occur in a sequence; if one repeat is detected one expects that more remain to be found. The lengths of repeats are usually between 20 and 50 amino acids, which is considerably shorter than most domains. An alignment including consecutive repeats should not be used for detection of outliers unless the number and distribution of repeats are absolutely conserved.

In the detection of repeats using SMART an algorithm is used that derives similarity thresholds that are dependent on the number of repeats already found in a protein sequence (Andrade *et al.*, 1999b). These thresholds are based on the assumption that suboptimal local alignment scores of a profile/HMM against a random sequence database are well described by an extreme value distribution (EVD). The result of this protocol is that acceptance thresholds for suboptimal alignments are lowered below the optimal scores of nonhomologous sequences.

Alignment scores generated from the comparison of a repeat profile with a database of randomized sequences are derived with Searchwise (Birney *et al.*, 1996), which uses a Smith–Waterman comparison (Smith and Waterman, 1981). A number n of score distributions for the 1st (optimal), 2nd (first suboptimal), and up to the nth highest scores of the profile compared with randomized sequences are fitted to n EVDs. Parameters are obtained for each fit that allow the transformation of alignment scores for the top n (sub)optimal alignments into E values. Since these E values are dependent on the repeat number, they are sensitive to the number of true-positive repeats in a sequence.

True-positive repeats are identified using two acceptance thresholds: a minimum E value and a minimum number of repeats required to occur in a sequence (e.g., *WD40* repeats are thought to occur in groups of at least six). These thresholds and the generation of an extensive alignment for a repeat family are defined manually after the method is applied to the current protein database.

Multiple alignments of repeats are constructed in an iterative manner. The initial alignment is based on definitions from determined protein structures or else from the literature. In the initial database search step, a profile constructed from the multiple alignment is compared with a sequence database. Top scoring sequences are considered using complementary approaches such as PSI-BLAST and FASTA to provide the two thresholds: minimum E value and minimum number of repeats per protein required. After one or two iterations, the final alignment and the thresholds are stored in the SMART database to allow the detection of repeats in any sequence.

3. *Sequence Motifs*

Highly conserved segments in proteins that are present outside of domains or else are incomplete portions of whole domains are termed motifs (Henikoff and Henikoff, 1991; Tatusov *et al.*, 1994; Bork and Gibson, 1996). Motifs may encompass active or binding site residues and, consequently, are frequently used to predict functional similarities between divergent homologs. Conserved families of sequences that are not folded in the absence of bound protein ligands are termed unstructured motifs. Examples of this phenomenon are the actin-binding motif of thymosin-β (**THY**), which has been shown to adopt a helical structure only when bound to actin (Van Troys *et al.*, 1996), and a staphylococcal protein, which is unfolded except when bound to mammalian fibronectin (Penkett *et al.*, 1998). A new example of a putative unstructured motif that arose out of a recent SMART update is a protein 4.1-binding motif (**4.1m**) in syndecans (Fig. 4, see Color insert; Table II).

The AT-hook (**ATh**) is an unusual example of a motif that is conserved in sequence and yet contains little secondary structure either in isolation or when bound to its ligand, DNA (Huth *et al.*, 1997; Aravind and Landsman, 1998). Additionally, sequence-similar motifs such as the helix–hairpin–helix motif (**HhH1, HhH2**) and "Asp-box" motifs, can occur within nonhomologous domain contexts (Doherty *et al.*, 1996; Russell, 1998). There is speculation that these arose in evolution either by gene duplication and insertion within a gene region coding for a separate domain, or by convergent evolution.

Sequence motifs are detected by SMART in a similar manner to domains. In situations where motifs are identified within detected domains, both the motif and the domain are shown.

II. Domain Families in Archaea, Bacteria, and Eukarya

A. *Horizontal Gene Transfer*

The burgeoning sequence data set, increasingly fed by the results of genome sequencing projects, affords an opportunity to assess the manner by which protein families have evolved. Before large-scale comparisons of complete genomes, the overwhelmingly predominant method of gene dispersal in cellular organisms was thought to be vertical transmission, through intragenome duplication and speciation. Thus, an intragenome duplication event would result in homologs that are termed "paralogs," and a speciation event would result in a pair of homologs that are termed "orthologs" (Fitch, 1970). Paralogs normally arise because of duplication of individual genes. They may also arise because of a whole genome duplication (polyploidy) (Ohno, 1970), of which there are predicted to have been at least two in the chordate lineage (Sidow, 1996), one in the *Saccharomyces cerevisiae* lineage (Wolfe and Shields, 1997) and several in ancestral plants (Gaut and Doebley, 1997).

The possibility that genes have been transferred horizontally between species, however, has long been mooted, in particular with respect to the origins of eukaryotic mitochondria (reviewed in Gray *et al.*, 1999). Thirty years ago, Margolis (1970) proposed an endosymbiotic origin of the mitochondrion based on the discovery of its separate genome, independent of that of the nucleus. Comparison of mitochondrial rRNA genes has suggested that the mitochondrial genome is monophyletic and that the likely evolutionary ancestor of the mitochondrion is related to the α division of modern Proteobacteria (Yang *et al.*, 1985). From the relatively small size of the mitochondrial genome it is assumed that the nuclear genome now contains many genes that have been transferred from the mitochondrial genome. Bacterial symbiont origins for eukaryotic plastids and other organelles are also indicated (McFadden *et al.*, 1994; reviewed in Corsaro *et al.*, 1999).

The results of comparative genomics studies indicate that the individual histories of protein families contain episodes of both vertical transfer and horizontal transfer of genes (Koonin *et al.*, 1997; Doolittle, 1998, Doolittle and Logsdon, 1998; Woese, 1998; Ponting *et al.*, 1999b). Inference of past horizontal transfer events depends on detecting significant differences between the topology of the phylogenetic tree for the gene family and that of the organismal tree. Large-scale horizontal gene transfers have been suggested between archaeal and bacterial lineages, and between bacterial lineages (e.g., Aravind *et al.*, 1998; Wolf *et al.*, 1999a; Nelson *et al.*, 1999). Such studies indicate that for ancient protein

TABLE II
Newly Identified Domain Homologs from Recent SMART Database Update

Domain/motif	Found in	Query (residues)	Method (iteration)	*E* value	Target sequence
4.1m	Syndecans	Bovine neurexin Iβ (381-437)	FASTA	1×10^{-2}	*Drosophila melanogaster* syndecan
C2 domain	MBC, CED-5, and DOCK180	Human KIAA0209 (412-582)	PSI-BLAST (1)	1×10^{-4}	C2 in *C. albicans* Vps34p.
DEP	p235 putative PI 5-kinase	Epac Rap1 GEF DEP (69-144)	PSI-BLAST (1)	3×10^{-5}	p235, phosphoinositide 5-kinase
ENTH	Sla2p, HIP1	Yeast Sla2p (1-258)	PSI-BLAST (2)	3×10^{-3}	ENTH in human epsin-2b
Fibrinogen-like domains	COLFI domains	*C. elegans* neurexin IV FBG-like domain (W03D8.6) (512-722)	PSI-BLAST (2)	2×10^{-4}	COLFI domain of chick collagen α_1(III)
GEL	Sec23p and Sec24p	Slime mold villin GEL (312-406)	PSI-BLAST (3)	6×10^{-4}	*A. thaliana* Sec23p (T7B11.7)
LamG (Jelly Roll fold)	Sialidases	Human agrin laminin G (1373-1509)	PSI-BLAST (1)	6×10^{-5}	*Streptomyces coelicolor* sialidase (gene SC4B5.07c)
LamG (Jelly Roll fold)	Usher syndrome type type IIa protein	*S. coelicolor* LamG domain (450-648)	PSI-BLAST (3)	2×10^{-4}	Human Usher syndrome type IIa

PH domain	IPL	Pleckstrin PH1 (1-105)	PSI-BLAST (4)	2×10^{-5}	Human *I*mprinted in *P*lacenta and *L*iver
RasGEF	BCAR3, HRSH2, Nsp1, Nsp3	BCAR3 (544-825)	PSI-BLAST (1)	2×10^{-7}	RasGEF in human Rap1 GEF, Epac
SH3 domain	MBC, CED-5, and DOCK180	*Drosophila* MBC (1-84)	PSI-BLAST (1)	1×10^{-3}	SH3 in human ArgBP2b
SH3 domain	Kakapo, plectin, and Bullous pemphigoid antigen 1	*Drosophila* Kakapo (917-971)	PSI-BLAST (3)	7×10^{-4}	Human ITK Tyrosine kinase SH3
VWA	Integrin β-subunits	Chick collagen α_1(VI) VWA (822-999)	PSI-BLAST (2)	8×10^{-4}	*Drosophila* integrin β subunit
VWA	Ku86/Ku70	Rat integrin α E2 VWA (193-380)	PSI-BLAST (7)	9×10^{-4}	Hamster Ku86
ZnF_AN1	AN1-type zinc finger	21 residue conserved alignment block	MoST (2)	2×10^{-2}	Hamster S mu bp-2
ZnF_AE	Archaeal/eukaryotic zinc finger	*A. fulgidus* AF0573 (1-54 [complete])	PSI-BLAST (0)	6×10^{-4}	*Drosophila* Pushover
ZnF_UBR1	*Drosophila* Pushover	*C. elegans* UBR1p (C32E8.11) (14-84)	PSI-BLAST (1)	7×10^{-4}	Zinc finger in yeast UBR1p
ZP	*C. elegans* cuticlins	*O.latipes* choriogenin H ZP domain (273-555)	PSI-BLAST (2)	6×10^{-4}	*C. elegans* cuticlin 1

families, complete congruencies between gene and organismal trees are rare, suggesting that cellular life is fundamentally of chimeric origin. Although acquisition of genes via horizontal transfer between eukaryotes is thought to be rare, the transfer of mobile elements or other parasitic sequences is less so (Kidwell, 1993) particularly in insects, although a LINE element was recently shown to be transferred from a snake to an ancestor of ruminants (Kordis and Gubensek, 1998).

A consequence of genome chimeras is that it is rare that one can accurately assign a particular protein family to a single phylogenetic lineage. Thus, assignments of domains as "prokaryotic-specific" or "vertebrate-specific" proteins, for example, are often inaccurate. Perhaps a more pertinent question is, in which lineage did the gene for the domain initially arise? Answering this conundrum requires considerable information on the gene family from phylogenetically diverse organisms and an assumption that vertical transmission of the domain has occurred more frequently than horizontal transfer. In addition, it raises the question of the genesis of domains. Since gene duplication appears to have been the major mechanism for the generation of domain families, the genesis of a domain can be defined as the genetic event that gave rise to a family of domain homologs that are not detectable as homologs of any other domain family. Thus our understanding of the origins of domains will alter as the methods of detecting homologs improve.

To illustrate the complexity of assigning the phylogenetic origin of domains, laminin G (**LamG**) domains, which arose from the recent SMART update (Table II), are analyzed. These domains are predicted to possess a jellyroll-type fold, based on significant sequence similarity to pentraxins (Beckmann *et al.*, 1998). In a PSI-BLAST search, domains with significant similarity to laminin G domains were found (Table II) in a *Streptomyces coelicolor* neuraminidase (sialidase; gene SC4B5.07c), *S. coelicolor* and *Saccharopolyspora rectivirgula* β-galactosidases (Inohara-Ochiai *et al.*, 1998), *Bacillus circulans* cycloinulo-oligosaccharide fructanotransferase (Kanai *et al.*, 1997), a *S. coelicolor* protein kinase (pkaG), human pregnancy-associated plasma protein A (Haaning *et al.*, 1996), a *S. coelicolor* putative protein (gene SC2H4.01), an integrin α and β4 homologue (Schwarz and Benzer, 1997; May and Ponting, 1999) in *Synechocystis* sp. (gene slr1028) and in human Usher syndrome type IIa protein. In the β-galactosidases this domain occurs as an insert within the catalytic domain. The laminin G-like (**LamGL**) domain encoded by the Usher syndrome type IIa gene occurs in its 5′ region. This region has not yet been found to be mutated in individuals with this sensorineural hearing deficiency and retinitis pigmentosa disorder (Eudy *et al.*, 1998).

The origins of laminin G domains are difficult to assess. The lack of detectable homologs in archaea argues for at least one horizontal gene transfer event between eukaryotes and bacteria. Yet, what of the direction of this transfer? On one hand, bacterial neuraminidases and the *Synechocystis* integrin α and $\beta 4$ homolog are predicted to contain domains that have been horizontally transferred from eukaryotes (Baumgartner *et al.*, 1998; May and Ponting, 1999), which suggests that the laminin G-like domains in these proteins also originated via horizontal transfer *from* eukaryotes. On the other hand, however, the jellyroll fold is known to be widespread in bacteria in hydrolases and toxins, which might indicate a bacterial origin, with subsequent horizontal transfer *into* eukaryotes. Indeed, these scenarios are equally parsimonious, and the possibility remains that horizontal transfers in *both* directions between bacteria and eukarya might have occurred.

B. Ancient Domain Families

Recent determinations of the complete genome sequences of organisms, in particular *Haemophilus influenzae* (Fleishmann *et al.*, 1995), *Methanococcus jannaschii* (Bult *et al.*, 1996), and *S. cerevisiae* (Goffeau *et al.*, 1996), have shown that many domain families are represented in each of the three forms of cellular life. Analysis using COGS (Tatusov *et al.*, 1997; Koonin *et al.*, 1998) shows that the majority of proteins possessing translation, ribosomal structure, and biogenesis functions, and some proteins involved in various metabolic processes (http://www.ncbi.nlm.nih.gov/cgibin/COG/readoganu?phy=ehugpcmy) are conserved in eight eukaryotic, bacterial, and archaeal genomes. However, these proteins represent only 13% of all COGS. The scarcity of ortholog conservation in these eight genomes contrasts with the finding that almost half of all protein folds are present in all three kingdoms of life (Wolf *et al.*, 1999b). This suggests that rapid mutation, duplication, deletion, and horizontal transfer events have radically reshaped these organisms' genomes from that of the hypothetical last common ancestor (the "cenancestor"), with relatively little remaining unchanged.

The conservation of orthologs, rather than paralogs, in each of the three forms of cellular life, is evidence for the preservation of function from the last common ancestor. However, it is known that nonhomologous proteins may possess essentially identical functions in different species (reviewed in Koonin *et al.*, 1996; Galperin *et al.*, 1998). It is proposed that such "nonorthologous displacement" of function occurs because of accumulative mutations within substrate-binding pockets or active sites. This might have been accelerated by large-scale horizontal gene transfer since this would increase the acquisition of beneficial mutations that result in novel function.

Comparative genomic analyses (Koonin *et al.*, 1997; Rivera *et al.*, 1998; Andrade *et al.*, 1999c) show that eukaryotic "informational genes" (those which function in translation, transcription, and replication) are most closely related to those of *M. jannaschii*, whereas "operational genes" (functioning in amino acid synthesis, biosynthesis of cofactors, fatty acid and phospholipid, the cell envelope, energy metabolism, intermediary metabolism, nucleotide biosynthesis, and regulatory functions) are more similar to those of bacteria. Apparent horizontal transfer of genes at such a scale has been interpreted as implying either a bacterial/eukaryotic chimera as the *M. jannaschii* ancestor (Koonin *et al.*, 1997), or else a bacterial/archaeal chimera as the earliest protoeukaryote (Rivera *et al.*, 1998).

An in-depth study of DNA repair systems (Aravind *et al.*, 1999a) has concluded that few, if any, repair proteins occur with identical collinear domain arrangements in all three kingdoms of life. Approximately 10 enzyme families of adenosine triphosphatases (ATPases), photolyases, helicases, and nucleases were identified that are all likely to have been present in the cenancestor. These enzymatic domains are accompanied in DNA repair proteins by numerous regulatory domains. This indicates that the domain architectures of these proteins are labile, with incremental addition and/or subtraction of domains to conserved cores to be a common phenomenon except in the most closely related species.

A second in-depth study, this time of domain families that function in eukaryotic signaling, showed that the great majority of enzymes (23 of 28 considered) possess homologs in prokaryotes (Ponting *et al.*, 1999b). Although some of these are thought to have arisen as a result of horizontal transfer from eukaryotes (see Ponting *et al.*, 1999b for details), there is evidence from their phyletic distributions that many were present in the cenancestor. The functions of many of these prokaryotic enzymes, however, are likely to be distinct from their eukaryotic counterparts. For example, Pkn2 from the bacteria *Myxococcus xanthus* is a protein serine/threonine kinase (**ST_Kc**) that is likely to regulate the activity of endogenous β-lactamase (Udo *et al.*, 1995), a phospholipase D (**PLD**) homolog is a bacterial endonuclease (Pohlman *et al.*, 1993), and bacterial clostripain and gingipain are cell-surface processing endopeptidases that are homologs of the apoptotic enzymes, caspases (**CASc**) (Chen *et al.*, 1998; Aravind *et al.*, 1999b).

By contrast few regulatory domains that function in eukaryotic signaling are detectable in prokaryotes (Ponting *et al.*, 1999b). Of the 185 domain/motif families studied, only nine occur in all three kingdoms of life. Of these, several are likely to have been disseminated by horizontal transfers, such as the cyclic nucleotide monophosphate binding domain (**cNMP**) from bacteria to *Archaeoglobus*, the polycystic kidney disease domain (**PKD**) between prokaryotes and eukaryotes, and fibronectin type III domains from eukaryotes to bacteria and archaea. However, the widespread occurrence of six domains and motifs indicates that these were present in the last common ancestor (cenancestor) of eukaryotes, archaea, and bacteria. These are cystathionine β-synthase (**CBS**) domains, a domain family exemplified by mammalian JAB (**JAB_MPN**), another exemplified by plant pathogenesis-related proteins of group 1 (**PR-1**), PSD-95, Dlg, ZO-1/2 (**PDZ**) domains, tetratrico peptide repeats (**TPRs**) and von Willebrand factor A (**VWA**) domains.

To understand general principles of protein evolution it is instructive to focus on specific examples. Here, VWA and other domains are discussed as representative families that are present in archaea, bacteria, and eukarya.

1. von Willebrand Factor A Domain Family

The finding of VWA domains in prokaryotes was unexpected, although it might have been anticipated since the VWA domain fold is commonly found in intracellular phosphoryl transfer enzymes. The newly identified VWA domain-containing proteins appear not to be restricted to extracellular localizations, and most are predicted to have retained the metal-binding sites observed in some eukaryotic extracellular homologs (Ponting *et al.*, 1999b). The domain architectures of prokaryotic VWA domain-containing proteins are dissimilar from those of eukaryotes, indicating that the domain family possesses multiple distinct functions. Ironically, although *Streptococcus pyogenes* and mammalian integrin $\alpha_5\beta_1$ VWA domain-containing proteins both bind fibronectin, the bacterial protein uses a separate region of its sequence to do so (Kreikemeyer *et al.*, 1995).

The VWA domains in some integrin α subunits are readily apparent from their sequences. Although much functional evidence (reviewed in Loftus and Liddington, 1997) supports a hypothesis that integrin β subunits also contain a VWA domain (Lee *et al.*, 1995; Bajt and Loftus, 1994; Tozer *et al.*, 1996; Tuckwell and Humphries, 1997), there has been no statistical evidence for significant sequence similarity.

However, the PSI-BLAST search method, using a very conservative inclusion threshold of $E < 10^{-4}$, can detect, with significance, the similarities in sequence between previously known VWA domains and integrin β subunits (Table II). The hypothesis that all integrin β subunits contain a VWA domain appears to be correct.

Similar searches detect VWA domains in the DNA-binding Ku70 and Ku80 proteins that are subunits of a heterodimeric autoantigen of approximately 70 and 80 kDa, respectively (Mimori and Hardin, 1986). Ku inhibits nucleotide excision repair by binding specifically to double-strand breaks and recruiting a large protein complex containing a DNA-dependent protein kinase (reviewed in Bertuch and Lundblad, 1998; Frit *et al.*, 1998). The VWA domain of Ku70 contains a region that has been proposed to participate in formation of the Ku70-Ku80 dimer (Wang *et al.*, 1998); hence the VWA domains of Ku70 and Ku80 might form a homotypic heterodimer. The VWA domains of integrin β subunits and Ku are now predicted by SMART.

2. *B7im/HC/Hf1K (Prohibitin) Domain Family*

E. coli Hf1C and Hf1K are homologous subunits of a dimeric complex that mediates homo-oligomerization of the membrane-associated protease FtsH (Hf1B) (Akiyama *et al.*, 1998). They are known to be homologs of human band 7 erythrocyte membrane protein (Noble *et al.*, 1993). However, we (SMART domain: **PHB**) and others (see PFAM domain **BAND_7** and COGS number 0330) have recognized that additional homologs, termed prohibitins, are present in eukaryotes. *S. cerevisiae* prohibitin 1 and 2 (Phb1p, Phb2p) are mitochondrial inner membrane proteins that form a Phb1p-Phb2p complex (Berger and Yaffe, 1998; Coates *et al.*, 1997) and regulate cellular replicative lifespan (Coates *et al.*, 1997). Thus it was predicted that prohibitins regulate cellular senescence by modifying the activities of mitochondrial FtsH-like enzymes. This prediction was borne out by recent studies that concluded that prohibitins regulate the proteolysis of membrane proteins by the Afg3p/Rca1p FtsH-like protease (Steglich *et al.*, 1999); the authors also noted that Hf1C, Hf1K, and prohibitins are homologs.

Prohibitin homologs are represented in each of the completely sequenced archaeal genomes. However, these organisms appear to lack FtsH orthologs. This argues for a function for these domains in archaea that is distinct from that of homologs in bacteria and in eukaryotic mitochondria. It is likely that **PHB** domains were present in the cenancestor, although FtsH-like molecules were not, and that FtsH-like molecules were introduced into the eukaryotic lineage from the protomitochondrion.

3. *Tail-Specific Protease Family*

The interphotoreceptor retinoid-binding protein (Borst *et al.*, 1989) functions in the regeneration of rhodopsin in the mammalian visual cycle. It is exclusive to vertebrates yet contains a repeated structure that has been found singly in bacterial and plant tail-specific proteases (**TSPc**) (Silber *et al.*, 1992) and the archaeal tricorn protease (Tamura *et al.*, 1996). The eukaryotic homologs of TSPc are likely to be inactive as proteases since they lack residues implicated in the active site of *E. coli* **TSPc** (Keiler and Sauer, 1995).

Sequence analysis implies that plant **TSPc** homologs appear to have been acquired from bacteria via horizontal transfer (results not shown). It is notable that no **TSPc** homologs have been observed in fungi or in invertebrates, even in the completely sequenced genome of *Caenorhabditis elegans*. The vertebrate homologs of this family, therefore, are likely to have arisen either via lineage-specific gene loss in fungi, invertebrates, and plants or, in a more parsimonious explanation, via horizontal transfer into the vertebrate lineage, probably from the bacteria. If the latter explanation gains greater credence, then the horizontal transfer of this gene into vertebrates would be seen to have contributed significantly to the evolution of the vertebrate eye.

4. *Two-Component Signaling Systems*

In bacteria and archaea, responses to environmental stimuli are elicited by so-called "two-component regulatory systems" of proteins with histidine kinase and/or receiver domains (Mizuno, 1998). Histidine kinases, which are members of a specific ATPase family (Mushegian *et al.*, 1997) (**HATPase**) mediate phosphotransfer to phosphoaccepting Che Y-like receiver (**REC**), and to histidine-containing phosphotransfer (**HPT**) domains. Recently, it has become apparent that regulation of these signaling events is complex, involving three additional families of domains: Per-Arnt-Sim (**PAS**) domains, which detect input signals and mediate dimerization events (Zhulin *et al.*, 1997; Ponting and Aravind, 1997; Taylor and Zhulin, 1999); **GAF** domains, which likely function in binding cyclic nucleotides (Aravind and Ponting, 1997); and, intracellular **HAMP** domains, which are likely to transmit conformational changes in transmembrane receptors (Aravind and Ponting, 1999).

It is striking that, although the cenancestor of cellular life is most likely to have contained similar signaling mechanisms, these systems have been almost completely superseded in the multicellular eukaryotes by protein kinases phosphorylating on serine, threonine, or tyrosine. The unicellular fungi *S. cerevisiae* and *Candida albicans* appear to have maintained or else appropriated through horizontal gene transfer a two-component signaling system (Posas *et al.*, 1996). The only kinases of the HATPase family remaining in multicellular eukaryotes, however, are plant phytochromes and ethylene receptors, probably acquired from an endosymbiont cyanobacterium, and pyruvate dehydrogenase kinase (Popov *et al.*, 1993). Similarly, of the **REC, HPT, PAS, GAF** and **HAMP** domain families, the only domains represented in metazoa are **GAF** domains in phosphodiesterases and **PAS** domains in numerous nonphosphorylation-dependent signaling pathways. It would appear that multicellular eukarya discarded much of the histidine kinase-mediated signaling machinery and evolved a separate and complex apparatus of signaling domains based on phosphorylation of Ser, Thr, and Tyr residues. The reason for this revolution in signaling remains unknown.

5. RNA-Binding Domains

A number of RNA-binding domains are identifiable in archaea, bacteria, and eukarya and consequently are likely to have been obligatory components of the cellular machinery since the existence of cells: the S1, S4, K homology, and PUA families of RNA-binding domains (**KH, PUA, S1, S4**) (Gibson *et al.*, 1993; Aravind and Koonin, 1999; Bycroft *et al.*, 1997) as well as the **HhH** motif (Doherty *et al.*, 1996) argued to bind RNA in some instances (Aravind *et al.*, 1999a). Other RNA-binding domains are found only in bacteria and eukarya, indicating possible acquisition by eukaryotes from the protomitochondrion: the R3H-type (**R3H**), double-stranded RNA-binding motif (**DSRM**) and cold shock protein (**CSP**) RNA-binding domains (although the similarity of the latter to the **S1** domain may argue for a more ancient heritage of these families, see Section I,B).

III. Domains Originating Early in Eukaryotic Lineage

A. *Horizontal Gene Transfer*

Several studies have concluded, from the isolated incidences of eukaryotic gene homologs in bacteria, that bacteria frequently have acquired eukaryotic genes by horizontal transfer. Domains likely to have originated in eukaryotic genomes, but observed in bacteria, are diverse in function and include β/γ crystallins, EF hands, and fibronectin type III, SET and SWIB domains, and leucine-rich and YWTD repeats (Swan *et al.*, 1989; Little *et al.*, 1994; Bagby *et al.*, 1994; Slack and Ruvkun, 1998; Stephens *et al.*, 1998; Ponting *et al.*, 1999b). By contrast BRCT and TIR domains are predicted, from the observation of divergent homologs in diverse bacteria but not in archaea, to have entered the eukaryotic lineage from bacteria (Aravind *et al.*, 1999a, 1999b).

The same direction, from bacteria to eukarya, was proposed for the horizontal transfer of SH3 domain homologous genes (Ponting *et al.*, 1999b). This proposal arose from the observation of a domain family in bacterial lytic proteins with significant similarity to mammalian SH3 domains (Whisstock and Lesk, 1999; Ponting *et al.*, 1999b). The direction of transfer was proposed to be into the eukaryotic lineage based on the observed lack of SH3 domain homologs in archaea and in plants. Hypothetical protein sequences from the plant *A. thaliana* that were recently deposited in databases (namely GeneBank: F19H22.120, T4L20.240 and T13E11.13), however, can be shown to contain obvious SH3 domains (data not shown). This indicates that all major branches of eukaryotic life contain this domain family and that horizontal transfer from eukarya to bacteria, with further propagation via horizontal transfer among bacteria, cannot be discounted. The three plant SH3 domain-containing sequences, however, appear to lack other signaling domains. Consequently, further investigation is required to determine whether these proteins function in intracellular signaling pathways.

A previously unrecognized example of likely horizontal gene transfer from eukaryotes to a bacterium relates to an animal arrestin-like homolog in *Bacillus subtilis.* Arrestins function by terminating G-protein-coupled receptors' activities upon binding, thereby abrogating interactions between receptors and G proteins. Monomeric arrestins contain a two-domain structure in which each domain is constructed from a seven β-strand sandwich (Hirsch *et al.*, 1999). Structural and sequence similarities between the two domains indicate that they are homologs. Hirsch *et al.* (1999) stated that the most distant homolog of visual arrestins occurs in invertebrates. However, PSI-BLAST analysis demonstrates that arrestins are members of an extended family of homologs that include numerous invertebrate and yeast representatives (Fig. 5, see color insert). Identified eukaryotic homologs include Rod1p, which influences drug tolerance in yeast (Wu *et al.*, 1996); yeast vacuolar sorting protein Pep8p (Bachhawat *et al.*, 1994); and a gene from the Down syndrome critical region of human chromosome 21q22.2 (Nakamura *et al.*, 1997). The existence of yeast arrestin homologs has been proposed previously (Chervitz *et al.*, 1998); however, alone among prokaryotic sequences a sporulation stage 0-control gene in *B. subtilis, spo0M* (Han *et al.*, 1998) was also identified as an arrestin homolog. Spo0M contains both arrestin domains, but is unlikely to possess similar functions to arrestins in *B. subtilis* given the lack of known G-protein-coupled receptors in bacteria.

By contrast to the many genes predicted to be of eukaryotic origin in bacteria, few instances of horizontal gene transfer from eukaryotes to archaea have been suggested (Makarova *et al.*, 1999). Only a single case of horizontal gene transfer from eukarya to archaea was predicted in the recent survey of eukaryotic signaling domains (Ponting *et al.*, 1999b). This was of a family of putative zinc fingers represented by the ubiquitin-like fusion protein AN1 (**ZnF_AN1**). (Recently detected eukaryotic members of this family include the DNA-binding protein S mu bp-2 [see

Table II].) However, during an analysis of the UBR1p family of zinc fingers (**ZnF_UBR**), which led to the identification of a new member of this family in the *Drosophila* pushover/calossin protein (Xu *et al.*, 1998) (Table II), a previously unidentified domain family was found, with members drawn only from archaea and eukarya (Fig. 6, see Color insert). This phyletic distribution suggests either that this domain originated in the last common ancestor of archaea and eukarya, or that the gene family has propagated between kingdoms via horizontal gene transfer.

B. Domain Families Represented in Fungi, Plants, and Metazoa

Table I shows that many domain families are widespread among fungi, plants, and metazoa and yet are absent from prokaryotes. It is assumed that these domains arose in early eukaryotes before the emergence of these three major eukaryotic lineages. Consideration of the known functions of these domains, and the proteins in which they occur, strongly suggests that emergence of several cellular functions that are unique to eukaryotes occurred in early eukaryotic history. These functions are likely to have coevolved with the abilities of the protoeukaryotic cell to reproduce sexually and to partake in cell–cell communication. Here we review several eukaryotic-specific domain families as illustrations of the coevolution of domain families with cellular functions.

1. Ubiquitin-Mediated Proteolysis Pathway

In eukaryotes, proteins are tagged for proteolytic degradation by the 26S proteasome by the attachment of multiubiquitin chains. Ubiquitination proceeds via the transferal of activated ubiquitin (**UBQ**) to a ubiquitin-conjugating enzyme (**UBCc**) usually in the presence of a ubiquitin ligase, E3. The ubiquitin ligase complexes contain proteins with domains involved in ubiquitin thioester intermediate formation (**HECTc**), domains acting as receptors for ubiquitin targets (**FBOX**) and domains that interact with UBCc proteins (**CULLIN**). All of these domains are absent from prokaryotes, as befitting organisms that lack this type of proteolysis pathway.

2. Apoptosis

The situation with the domain families of the ubiquitin-mediated pathway contrasts with the domain families that function in animal and plant programmed cell death, or apoptosis. As reviewed elsewhere (Aravind *et al.*, 1999b), a few domains in eukaryotic apoptotic proteins have prokaryotic homologs, including the cysteine protease family of caspases

(**CASc**), and the Toll-interleukin resistance (**TIR**) domain family. It is significant that fungi lack many of these apoptotic domains (with the exceptions of **BIR** and **MATH** domains) since many of the morphological effects associated with animal and plant apoptosis have not been observed in yeasts (Fraser and James, 1998). Possible conservation of some features of apoptosis that are linked with the ubiquitin pathway have been suggested by the observation of putative MATH domain-containing ubiquitin hydrolases in yeasts and animals (Aravind *et al.*, 1999b). In addition, the cell death-related engulfment gene family CED-5/DOCK180/MBC is represented in yeast (Wu and Horvitz, 1998), indicating that this cellular function is widespread in all eukaryotes (cf. Table II). This family can be shown shown to contain single **C2** and **SH3** domains (Table II) indicating that polyproline-binding to **SH3** domains, and phospholipid-binding to **C2** domains, are involved in the function of these proteins during cell-corpse engulfment.

3. Phosphorylation and Second Messenger-Mediated Signaling Pathways

Phosphorylation of serine, threonine, or tyrosine residues by protein kinases, and their dephosphorylation by protein phosphatases, are critical mechanisms by which information-relaying signals are transduced in eukaryotic cells. Although protein kinases are by no means an eukaryotic invention (see Leonard *et al.*, 1998 for details), the large numbers of protein kinases in eukaryotes (118 in *S. cerevisiae* and 435 in *C. elegans* (Chervitz *et al.*, 1998)) reflect their importance in a multitude of diverse cellular processes. Eukaryotes have evolved signaling pathways that exploit the dual state of an amino acid, dependent on its state of phosphorylation, both as a signaling mechanism and as a means of colocalization of molecules within multimolecular complexes.

The best studies of signaling pathways are the mitogen-activated protein (MAP) kinase pathways of budding yeast (reviewed in Widmann *et al.*, 1999). These pathways contain a three component module: a MAP kinase, which is a substrate for a MAP kinase kinase, that in turn is a substrate for a MAP kinase kinase kinase. Although these modules are relatively well conserved across all eukaryotes, the number of MAP kinase modules, the identity of the pathway's initiating stimulus, and the cellular response to the signal are variable among diverse eukaryotes. In particular, the regulatory proteins that interact with the conserved MAP kinase modules are mostly not identical in domain architectures when compared between different species.

In yeast, the MAP kinase Fus3 induces cell cycle arrest via the degradation of cyclins, Cln1 and Cln2. The mitotic cyclins (**CYCLIN**) are cell cycle proteins that bind the protein kinase Cdc2 during interphase

(Murray and Hunt, 1993). Cyclins, kinases, and phosphatases that regulate the passage of the cell through the $G_1 \rightarrow S$ phase transition are all present in mammals, invertebrates, and plants (Solomon, 1993; Doonan and Fobart, 1997; Zavitz and Zipursky, 1997). However, multicellular eukaryotes contain multiple orthologs of yeast cell cycle proteins; they initiate proliferation via growth factors, rather than, for example, yeast mating factors, and they possess additional checkpoint controls and repair pathways.

Evolution of these signaling pathways has generated several domain families with members that bind phosphoserine- or phosphothreonine-containing proteins (**14-3-3** and **WW** domains), or phosphotyrosine-containing proteins (**PTB, PTBI** and **SH2** domains). A possible addition to this list are forkhead-associated (**FHA**) domains, which, in at least one case (Sun *et al.*, 1998), bind protein in a phosphorylation-dependent manner. However, FHA domains are not specific to eukaryotes, and it is suggested that they and PKN-2 protein kinases have undergone coordinated horizontal gene transfer among the bacteria (Ponting *et al.*, 1999b). Somewhat surprisingly, given that tyrosine-specific protein kinases in yeast are well established (Schieven *et al.*, 1986), *S. cerevisiae* appears to contain none of the flavors of phosphotyrosine-binding domains, except for a single SH2 domain in the nuclear protein Spt6p (Maclennan and Shaw, 1993). Thus, the extended families of protein tyrosine kinase- and SH2 domain-containing proteins are metazoan inventions (Hunter and Plowman, 1997).

Lipid products of phospholipases, DAG kinase, and phosphoinositide 3-kinase have also been recruited to the signaling cause, early in eukaryotic history. Fungi, plants, and animals have considerable numbers of lipid-binding signaling domains. Among these are DAG-binding (**C1**), phosphatidylserine-binding (**C2**), phosphoinositide-3-phosphate (PI(3)P)-binding (**FYVE**), and PI(3,4)P2- and PI(3,4,5)P3-binding (**PH**) domains that appear to have arisen early in the eukaryotic lineage. There are several apparently eukaryotic-specific signaling domains that adopt the PH domain fold. These include the Ran-binding domain (**RanBD**), the EVH1/WH1 (**WH1**) domain, and two flavors of phosphotyrosine binding domains (**PTB, PTBI**) (Prehoda *et al.*, 1999). Currently this fold is specific for domains involved in signaling and these families occur only in eukaryotes. Thus it is tempting to speculate that these sequence families all arose from an early eukaryotic common ancestor. The apparently rapid sequence divergence of these families and their multiple ligand-binding modes (**PH** domains bind phospholipids and proteins, **PTB** domains bind phospholipids and phosphotyrosine-containing polypeptides, and **WH1** domains bind polyproline-containing polypeptides) would be consistent with this proposal.

4. GTPase-Mediated Signaling Pathways

The origin of the family of Ras-like small GTPases, like many other enzyme families, is thought to predate the emergence of eukaryotes since a separate subfamily of small GTPases is present among the archaea and a subset of bacteria (Ponting *et al.*, 1999b). Although, as stated previously, the functions of prokaryotic proteins are often distinct from their eukaryotic homologs, there is a report of a eukaryotic small GTPase, yeast Sar1p, complementing the function of a bacterial ARF-like homolog in a *M. xanthus* knockout strain (Hartzell, 1997).

The family of eukaryotic Ras-like small GTPases may be divided into subfamilies, namely those of ARF, Rab, Ran, Ras, Rho, and Sar (**ARF, RAB, RHO, RAS, RHO, SAR**), which all contain representatives from fungi, plants, and metazoa. Consequently, these subfamilies and their cellular functions are likely to have emerged early in eukaryotic history. This implies that the last common ancestor of fungi, plants, and metazoa possessed vesicular transport (ARF and Sar), membrane trafficking (Rab), nuclear transport (Ran), signal transduction (Ras), and regulation of the actin cytoskeleton (Rho) functions.

Similarly, heterotrimeric G proteins are ubiquitous in eukarya, and the signaling pathways in which they participate are presumed to have evolved in a primitive eukaryote. Gγ subunits of G proteins (**GGL**) are likely to be motifs that are unstructured except in the presence of Gβ (Snow *et al.*, 1998). G protein β subunits are WD40 repeat-containing β-propeller structures. WD40 domains are presumed to have evolved from the many bacterial proteins with β-propeller structure (Murzin, 1992). However, aside from cyanobacterial homologus, which are clear examples of horizontal transfer from eukaryotes (Ponting *et al.*, 1999b), there has been little sequence-based evidence for this proposal until recently. Bacterial TolB protein sequences have been shown to possess statistically significant similarities to WD40 proteins (Ponting and Pallen, 1999), indicating that the latter are relatively ancient in origin. G protein α subunits are GTPases that are clearly related to Ras and to prokaryotic enzymes. The proliferation of α subunits' numbers, relative to those of β and μ subunits, in metazoa is clearly linked to the requirements of multiple organism-specific signaling pathways (Jansen *et al.*, 1999).

These GTPases cycle between inactive GDP-bound forms and active GTP-bound forms. Eukaryotic-specific domain families have evolved that either promote GTPase activities (GTPase activator proteins, "GAPs") or promote exchange of GDP for GTP (guanine nucleotide exchange factors, "GEFs"). Each of the Ras-like small GTPase subfamilies can be linked with a corresponding GAP family and a GEF family. The high-resolution structures of many of these GAPs and GEFs have now been determined, showing that GAPs specific for (some) members of the Ras subfamily (**RasGAP**) are likely to be distant homologs of GAPs specific for (some) members of the Rho subfamily (**RhoGAP**) (Scheffzek *et al.*, 1998 and references therein). However, the remaining GAP and GEF families do not appear to be structurally and evolutionarily related.

Although the origins of these GAPs and GEFs lie close to the base of the eukaryotic phylogenetic tree, the proteins in which they occur are more recent inventions. It is striking that of the 35 known yeast GAP and GEF proteins specific for Ras, Rho or Arf, only 7 are predicted by SMART to contain a multidomain architecture that is shared with a putative *C. elegans* ortholog (namely Bud2p/Cla2p, Lte1p, Bud5p, Scd25p, YBR260c, YBL060w, and SYT1). By contrast, the majority of worm GAP- or GEF-containing proteins have one or more orthologs in mammals with identical domain architectures. Similarly, it is expected that completion of the genome of *A. thaliana* will show that this plant contains GAP and GEF-containing proteins that are mostly dissimilar in modular architectures to those of yeast and those of metazoa. This situation is similar to the kinases: of 118 *S. cerevisiae* protein kinases only 2 possess putative orthologs in *C. elegans* (namely, Vps15p and Dun1p).

5. *Cytoskeleton*

Evolution of both the actin-based and the microtubule-based cytoskeleton have drawn on ATPases and GTPases that are likely to have been present in the cenancestor. The eukaryotic-specific molecules actin and tubulin β/γ polymerize to form filaments that form the basis of the cytoskeleton's structural integrity. Eukaryotic actins are members of a large family of ATPase homologs that also includes bacterial sugar kinases and heat shock proteins (Bork *et al.*, 1992). Eukaryotic tubulin β and γ subunits are GTPases that are homologs of bacterial FtsZ (Mukherjee and Lutkenhaus, 1994) as further demonstrated by their high resolution structures (Nogales *et al.*, 1998; Löwe and Amos, 1998). In addition, the molecular motors that translate across the cytoskeleton are also homologs of ancient enzymes. Myosins, kinesins, and zyneins are ATPases that possess structural features common among themselves and among wider families of ATPases (Kull *et al.*, 1996; Neuwald *et al.*, 1999).

Although the building blocks of the eukaryotic cytoskeleton appear to be ancient, the protein domains interacting with it appear to have emerged more recently. Several actin-binding domain families, namely calponin homology, CH, actin depolymerisation factor (**ADF**), the Sla2p

C terminus (**ILWEQ**), WASp homology 2 (**WH2**), profilin (**PROF**), and cyclase-associated protein, domains are all present in fungi, plants, and metazoa. Many of these domains bind similar sites on actin, although they possess different properties with respect to actin polymerization (reviewed in Van Troys *et al.*, 1999).

Although the gelsolin family of actin-binding domains **GEL** was thought to be present throughout the eukarya except in fungi (Schleicher *et al.*, 1988), we have identified (Table II) gelsolin homology domains at the C termini of yeast, plant, and metazoan Sec23p and Sec24p proteins. These proteins are constituents of the coat protein complex II (COPII) that generates secretory vesicles at the endoplasmic reticulum (Pagano *et al.*, 1999). These vesicles contain secretory proteins and travel from the endoplasmic reticulum to the Golgi apparatus. The finding of a **GEL** domain in the COPII proteins, Sec23p and Sec24p, implies that these regions mediate the interaction of the vesicle with the actin cytoskeleton.

Thymosin-β and villin headpiece actin-binding motifs (**THY, VHP**) are proposed to bind actin in a similar manner via an α helix succeeded by a 'Leu-Lys-Lys' motif (Van Troys *et al.*, 1999). These sequence characteristics are also prominent in WH2 motifs (Gertler *et al.*, 1996). It would appear that these motifs contain a smilar arrangement of α helices, as seen in the villin headpiece structure (McKnight *et al.*, 1997) in order to interact with actin. In HMMER2 searches using these motifs and an E value threshold of 0.1, we have been able to identify similar motifs in eukaryotic cyclase-associated proteins and nucleopolyhedroviral proteins (Fig. 7, see Color insert). It is suggested that these motifs possess actin-binding functions. The viral proteins might function in recruiting the host–cell actin cytoskeleton to move from the cytoplasm to the cell surface (cf. Cudmore *et al.*, 1995).

Another family that is present throughout eukaryotes and is involved in maintenance of the cytoskeleton is the *E*psin *N-t*erminal *h*omology (**ENTH**) domain family (Kay *et al.*, 1999). A previously-unidentified **ENTH** domain was found (Table II) in *S. cerevisiae* Sla2p (also known as End4p, Mop2p). This observation is consistent with previously described ENTH domains since the Sla2p ENTH domain is known to be required for endocytosis and actin organization (Wesp *et al.*, 1997). Huntingtin interacting proteins, which are mammalian homologs of yeast Sla2p (Kalchman *et al.*, 1997; Wanker *et al.*, 1997), also posses the ENTH domain. This suggests that the normal function of the Huntington disease gene product, huntingtin, might be related to endocytosis.

Many cytoskeletal and other metazoan proteins that are absent in yeast contain domains that are present in yeast. Thus it would appear that existing domains are "reused" in contrasting contexts during the evolution of individual eukaryotic lineages. For example, the animal paralogs dystrophin and utrophin, which function in maintenance of the neuromuscular junction, and their single ortholog in invertebrates contain **CH**-type actin-binding domains, a **WW** domain and a ZZ zinc finger (**ZnF_ZZ**) (Castresana and Saraste, 1995; Bork and Sudol, 1994; Ponting *et al.*, 1996). Yeast WW domain homologs function as splicing factors (Ess1p and Prp40p) and in the ubiquitin-mediated proteolysis pathway (Rsp5p), whereas a yeast ZZ domain occurs in a transcription factor (Ada2p). Thus, different eukaryotic organisms have made use of WW and ZZ domains for completely different cellular functions.

6. *Extracellular Proteins*

The greatest variations in protein and domain complements for different eukaryotic organisms are observed for extracellular proteins (Chervitz *et al.*, 1998; Copley *et al.*, 1999). Extracellular domain families that are apparently lacking in fungi include growth factor domains (**IIGF, NGF, TGFB**), interleukins (**INTERLEUKIN_2, INTERLEUKIN_4_13, INTERLEUKIN_10**), protease inhibitors (**SERPIN, KAZAL, KUNITZ, TIMP**), domains that frequently occur in metazoan extracellular proteases or transmembrane receptors (**APPLE, KR, CCP, CLECT, CUB, FU, GLA, LINK, TNFR, TSP1**), and domains that occur in extracellular matrix proteins (**C4, COLFI, FBG, FN1, FN2**) (Table I).

However, not all metazoan extracellular domains are missing in fungi. Epidermal growth factor-like (**EGF**) (Hogan *et al.*, 1995), low-density-lipoprotein receptor class A (**LDLa**) (De Virgilio *et al.*, 1996; Copley *et al.*, 1999), Lysin motif (**LysM**) (Birkeland, 1994; Ponting *et al.*, 1999b), WSC (Verna *et al.*, 1997; Ponting *et al.*, 1999c), and chitin-binding (**ChtBD**) (Butler *et al.*, 1991) domain families are all represented in metazoa and fungi. In addition, fibronectin type III (**FN3**), von Willebrand factor domain A (**VWA**) and pathogenesis related 1 (**SCP**) domains are present both in metazoan *extracellular* proteins, and in fungal, metazoan, and prokaryotic *intracellular* proteins (Ponting *et al.*, 1999b).

Vertebrates contain several proteins that maintain the integrity of the blood plasma circulatory system. These contain domains that are specific to vertebrates (**Gla, FN1, FN2**) (Patthy, 1985), domains that are found in different contexts in invertebrates and/or protists (**FBG, APPLE, KR**) (Xu and Doolittle, 1990; Eschenbacher *et al.*, 1993; Wilson *et al.*, 1993) and a domain that is found in all cellular life (trypsin-like serine protease, **Tryp_SPc**). The invertebrate versions of these domains, however, are found in molecular contexts that differ considerably from their vertebrate extracellular counterparts, indicating that although these nonenzymatic domains are likely to have arisen early in metazoan evolution, as might be expected, the proteins of blood coagulation and fibrinolysis are vertebrate inventions.

Fibrinogen and collagen appear to be inventions of early metazoan life (Xu and Doolittle, 1990; Exposito and Garrone, 1990). Although they were not previously thought to be homologs, PSI-BLAST searches reveal significant similarities between fibrinogen-like domains (**FBG**) and the C-terminal domains of fibrillar collagens (**COLFI**). It is suggested that these domain families share an early metazoan ancestor (Fig. 8, see Color insert). Although these domains could not be accurately aligned throughout, comparison with the known crystal structure of fibrinogen fragment D (Spraggon *et al.*, 1997) suggests that they adopt the same fold.

7. *Chromatin Remodeling*

Many of the factors that mediate chromatin remodeling appear to have evolved early in eukaryotic history. SWI-SNF-like complexes have been identified in yeast, plants, and metazoa (Côté *et al.*, 1994; Imbalzano *et al.*, 1994; Brzeski *et al.*, 1999; Jeddeloh *et al.*, 1999) and contain proteins with domain families that are peculiar to eukaryotic life. These domain families are bromo domains (**BROMO**) with histone H4-binding functions (Ornaghi *et al.*, 1999), "bromo-adjacent homology" domains (**BAH**) with protein-binding functions (Callebaut *et al.*, 1999), chromo (**CHROMO**) and chromo shadow (**ChSh**) domains with homodimerisation properties (Cavalli and Paro, 1998; Yamada *et al.*, 1999), and **PHD** and **SANT** DNA-binding domains (Aasland *et al.*, 1995; Aasland *et al.*, 1996). Two other domains of unknown function, **SET** and **SWIB,** are found in eukaryotic chromatin remodeling proteins and also in two *Chlamydia* proteins that are likely to have arisen via horizontal transfer from a eukaryotic source (Stephens *et al.*, 1998).

However, the packing of DNA into nucleosome-like structures is not unique to eukarya; similar structures appear in archaea (reviewed in Reeve *et al.*, 1997). Additionally, histones and minichromosome maintenance proteins (**MCM**) are widespread among eukarya and archaea and absent in prokarya, and the eukaryotic chromo domain has a structure that is highly reminiscent of archaeal histones that are involved in formation of archaeal chromatin (Ball *et al.*, 1997). Consequently, it is possible that chromatin remodeling in eukaryotes is an elaboration of a similar cellular mechanism in archaea.

Surprisingly, *C. elegans* appears to have lost a considerable number of chromatin proteins from the Polycomb group of proteins, observed in *Drosophila* and in mammals, although other transcription factor genes are mostly retained (Ruvkun and Hobert, 1998). This loss has been suggested to be associated with the observed dispersal of homeobox gene clusters (Ruvkun and Hobert, 1998). Interestingly, those Polycomb genes that are observed in *C. elegans* are exactly those that have been observed in *Arabidopsis* (reviewed in Preuss, 1999). It will be interesting to observe, on completion of the *Arabidopsis* genome project, whether these genes represent the core set necessary for chromatin remodeling in eukaryotic life.

IV. Domain Families in Multicellular Organisms

From what is known from the complete *C. elegans* genome, the evolution of multicellularity in eukaryotes appears to have required considerable genesis and expansion of domain families (Chervitz *et al.*, 1998; Copley *et al.*, 1999; Ponting *et al.*, 1999b). Domain genesis appears to have been most prevalent among extracellular domains (see Section III,B,6), whereas expansion of preexisting domain families, such as the well-known example of **PDZ** domains, appears to have occurred more frequently for intracellular domains (Chervitz *et al.*, 1998; Copley *et al.*, 1999). Expansions of families in vertebrates are likely to have been assisted by two independent genome duplications thought to have occurred in the chordate lineage (Sidow, 1996). On the other hand, as completely sequenced eukaryotic genomes become more numerous, it is likely that lineage-specific gene deletion will be seen as an important factor in genome evolution. The *C. elegans* genome, for example, appears to lack representatives of hedgehog, Toll/IL1 and JAK/STAT pathways (Ruvkun and Hobert, 1998).

A. *Domain Genesis*

Comparison of the complete genomes of *C. elegans* and *S. cerevisiae* and the incomplete genomes of *A. thaliana* and *H. sapiens* demonstrates the presence of several domain families that occur in only one of these lineages. For example, Mbp1p-like and GAL4-like (**GAL4**) DNA-binding domains occur only in fungi, and Bowman–Birk and squash-type protease inhibitors (**BowB, PTI**) are known only in higher plants. Vertebrates contain large numbers of well-characterized domains not found elsewhere. These include apoptotic domains (**CARD, DEATH, DED**) and hormones (e.g., **GHA, GHB**) and a hormone receptor domain (**HormR**). The full extent of these lineage-specific families will soon become apparent after completion of the human and plant genome sequencing projects.

C. elegans contains a large number of genes that appear to be nematode-specific (Chervitz *et al.*, 1998; Blaxter, 1998). Of these, some contain domains that have not been detected with significance other than in nematodes. For example, extracellular domains of the "Worm-specific repeat 1" (**WR1**) family occur in more than 200 copies in 34 *C. elegans* proteins, including several proteins with interspersed **KU** and **WR1** domains (e.g., Y43F8B.3) and a receptor kinase (D1044.3). Another domain is the "Worm-specific N-terminal domain" (**WSN**), which often occurs at the N termini of intracellular proteins containing, for example, **BRCT** and **ANK** repeats (e.g., F37A4.4 and F40E12.2) or protein tyrosine phosphatase domains (e.g., W03F11.4 and R155.2). It is not expected that the **WR1** and **WSN** domain families represent novel folds, but instead are likely to form subfamilies of larger sets of homologs. Indeed, the **WR1** domain shows many characteristics of the **EGF** domain family and may represent a divergent **EGF** homolog.

B. *Expansion of Domain Families*

The expansion of a domain family within a single lineage is likely to represent an evolutionary response to specific selection pressures. Examples of this phenomenon occur in all forms of cellular life. Higher plants contain a large multigene family of receptor protein kinases that are involved in development and pathogen resistance (Satterlee and Sussman, 1998). *Synechocystis* sp. PCC6803 has a larger set of two-component signaling systems than expected from its genome size. This might reflect special environmental sensing requirements for this photoautotrophic organism. *C. elegans* has a large repertoire of channels and receptors that mediates its neural system (Bargmann, 1998). It also contains expanded sets of nuclear hormone receptors (Sluder *et al.*, 1999), receptor tyrosine kinases (Ruvkun and Hobart, 1998), and proteins with one or more ShK toxin-like domains (**ShKT**) (Copley *et al.*, 1999) for less well-understood reasons.

A domain family that is considerably expanded in nematodes, relative to vertebrates, is the zona pellucida (**ZP**) domain (Bork and Sander, 1992). In database searches this domain was found in *C. elegans* cuticlin-1 (cut-1), a component of the nematode cuticle (Sebastiano *et al.*, 1991), and 33 other *C. elegans* proteins (Table II). On the basis of disulfide-linked domains that accompany the **ZP** domain in these proteins, it is likely that they localize to the worm's extracellular matrix. Indeed, it is possible that most of these proteins are components of the worm cuticle. The cuticle structure is the multilayered elastic exoskeleton that determines the worm's body shape. Although vertebrates lack an equivalent structure, the vertebrate egg envelope possesses many of the characteristics of the worm cuticle. This envelope, or zona pellucida, is an elastic outer layer of the ovum that contains sperm receptors. The sperm receptors and the invertebrate cut-1-like homologs are notable in both containing **ZP** domains. This further emphasizes the similarities, and potential homology, between the vertebrate zona pellucida and worm cuticle structures.

V. Domains in Diverse Molecular Contexts

A. *Genetic Mobility*

The frequency of lineage-specific proliferation of domain families suggests that genes encoding novel domain combinations can be generated by the shuffling of preexisting genes (Gilbert, 1978). Retrotransposition of long interspersed nuclear elements (Moran *et al.*, 1999) might account for the genesis of recently duplicated eukaryotic genes via exon shuffling, such as those encoding extracellular proteins (Patthy, 1996). However, it has been argued that there is little evidence for the participation of exon shuffling processes in the genesis of more ancient genes, such as those that first arose in early eukaryotes (Bork, 1996).

Many domain types demonstrate a strong propensity to occur as repeats within a single polypeptide. Such repetition of domains results initially in functional degeneracy, although this may be ameliorated in time by the divergence of the repeats' sequences, leading to functional divergence. For example, the human hypothetical protein KIAA0782 contains 5 **PH** domains. Given that **PH** domains are known to bind several phosphoinositides and several proteins (Shaw, 1996), it is predicted that these five domains possess different specificities for diverse ligands. However, repeats may possess synergistic functions for the multidomain protein. First, repeats may be required for the adoption of a stable tertary structure, such as for β-propellers. Second, tandem domains may possess affinities for similar ligands, thereby functioning in clustering multiple ligands, such as for **PDZ** domain-containing proteins (reviewed in Ponting *et al.*, 1997). Third, tandem domains may bind a single ligand with higher affinity compared with a single repeat, such as for the actin-binding **CH** domains (Gimona and Winder, 1998).

Although many typically extracellular domains are entirely absent from intracellular proteins, and vice versa, there is no absolute partitioning of domain families into separate cellular localizations. Several domain families, such as **VWA** (see Section II,B,1), **PDZ** (Wu *et al.*, 1999), **C2**

(Ponting and Parker, 1996), annexin II (Chung and Erickson, 1994), and actin-binding **GEL** (Wen *et al.*, 1996) domains have both intracellular and secreted members. An intracellular homolog of the extracellular plant bulb-type mannose-binding lectin domains (**B_lectin**) is present in *Dictyostelium discoideum* (Jung *et al.*, 1996). This bulb-type lectin-containing protein, termed comitin, is not only unusual in being intracellular, but it contains none of the disulfide bridges that characterize the plant bulb-type lectin structure (Hester *et al.*, 1995). Comitin appears to share a mannose-binding function with its plant homologs, yet unusually it also is known to bind actin (Jung *et al.*, 1996). Comitin is also exceptional in being the only bulb-type lectin homolog known outside of plants, suggesting that it was acquired by *Dictyostelium* from plants via horizontal gene transfer.

B. Domain–Domain Correlations

Although domains are often mobile and occur in many different modular architectures, it is notable that the co-occurrence of domains within single polypeptides is far from random, since a domain is usually found to co-occur only with a small subset of all domain types. When two domain types are not observed within the same molecule, it is likely that their activities are antagonistic, thereby effectively neutralizing the overall function of the molecule. Such an example is provided by protein kinase and phosphatase domains that are not currently known to co-occur within the same molecule. However, the reasons that functionally distinct and otherwise widespread domains have never yet been found together, such as signaling **PDZ** and **SH2** domains, remains elusive.

An example of the correlated co-occurrence of domains is exemplified by the **SH2** domain family. This domain is combined with only 15 other domain types in *C. elegans*. This is a relatively small number given that this organism possesses more than 100 different domains that function in intracellular signaling. The rate of domain combination within multi-domain proteins appears to be higher in vertebrates than in invertebrates, since approximately twice (27) the number of domains are currently found with SH2 domains in human protein sequences than in worm sequences. However, these figures demonstrate that most domains co-occur with relatively few of the total number of sequence families, given that such families number in the thousands. A consequence of this is that ill-characterized domain families may be predicted to possess a particular cellular function simply on the basis of co-occurring domains. For example, the function of **PX** domains (Ponting, 1996) remains unknown, yet its presence in proteins with well-described signaling domains argues for its participation in signal transduction processes.

In addition to this classification of cellular function by domain co-occurrence, analyses of domain combinations can also be used to improve the prediction of a protein's function. The **RhoGEF** domain, for example, is invariably found N-terminally to a **PH** domain. The co-occurrence of these two domains appears to be correlated with altered electrostatic potential, thereby resulting in prevention of the **PH** domain from binding phospholipids (Blomberg *et al.,* 1999). As this is a frequent function of the **PH** domain, the determination of a protein's domain architecture can assist in discounting a specific predicted function.

There is little doubt that a major cause of the partitioning of domains into functionally related co-occurring clusters relates to the compartmentalization of function inside and outside of cells. For example, the fusion of an intracellular domain to an extracellular domain might be selected against owing to an aberrant localization of function. Indeed, this is proposed to be responsible for oncogenic kinase activation leading to generation of a papillary thyroid carcinoma (Butti *et al.,* 1995; Greco *et al.,* 1993). In this example, the carcinoma is associated with a chromosomal rearrangement that results in replacement of the extracellular domain of the neurotrophic tyrosine kinase receptor by part of the intracellular tropomyosin-3. Even the combination of domains with similar functions, such as nucleotide binding, might be lethal. A Ewing's sarcoma, for example, is associated with the replacement of a RNA-binding **RRM** domain by a DNA-binding **ETS** domain (Jeon *et al.,* 1995; Peter *et al.,* 1997).

A variety of domain or motif families occur only as extensions to other domains. The Bruton's tyrosine kinase motif (**BTK**), for example, is found only at the C terminus of PH domains. Similarly, a C-terminal extension (the **S_TK_X** domain) to some subfamilies of serine/threonine kinases (**S_TK**) is not found in isolation. Cases where only the extension, and not the preceding domain, is found are strong evidence that the proteins are wrongly assembled from genomic sequence or else represent partial cDNA sequences (Fig. 9, see Color insert). Indeed, all five proteins annotated in SMART as containing a **S_TK_X** domain with no catalytic domain are noted to be fragments in their corresponding sequence database entries.

Correlations in the co-occurrence of domains can assist in the identification of distant members of a protein family that are not detected with significance using standard database searching methods. In all known examples of proteins with **C1** and **CNH** domains, for example, there is an intervening **PH** domain (Schultz *et al.,* 1998). The only exception to this rule is *C. elegans,* a hypothetical protein K08B12.5 (Fig. 9). Performing a database search with this intervening sequence yields other proteins with identical domain organization, but only at *E* values of 1 are other PH domain sequences detected. Thus only a comparison of this sequence to the similar domain architectures of other proteins results in the correct prediction of a PH domain for this sequence.

VI. Conclusions

Considerable advances have been made in the detection of homologs on the basis of significant sequence similarity. These methods, however, cannot be applied directly to the understanding of protein evolution and function. For this understanding to occur, it is informative to decompose proteins into their component domains using recently established domain database tools. Consideration of such domain architectures allows studies of the phyletic distributions of domains that assist in predicting the evolution of function. It is clear that representatives of a single domain family often possess distinct functions. Consequently, investigations are required to define the diversity of functions represented by single families using domain correlations, annotation of functional motifs, and mining of known three-dimensional protein structures. The successful use of these approaches and their reflection in the annotation of the widely used sequence databases are an essential prerequisite to the prediction of multimolecular pathways and complexes.

References

Aasland, R., Gibson, T. J., and Stewart, A. F. (1995). *Trends Biochem. Sci.* **20,** 56–59.

Aasland, R., Stewart, A. F., and Gibson, T. (1996). *Trends Biochem. Sci.* **21,** 87–88.

Akiyama, Y., Kihara, A., Mori, H., Ogura, T., and Ito, K. (1998). *J. Biol. Chem.* **273,** 22326–22333.

Allison, T. J., Wood, T. C., Briercheck, D. M., Rastinejad, F., Richardson, J.P., and Rule, G. S. (1998). *Nature Struct. Biol.* **5,** 352–356.

Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D. J. (1997). *Nucleic Acids Res.* **25,** 3389–3402.

Altschul, S. F., and Koonin, E. V. (1998). *Trends Biochem. Sci.* **23,** 444–447.

Andrade, M. A., Brown, N. P., Leroy, C., Hoersch, S., de Daruvar, A., Reich, C., Franchini, A., Tamames, J., Valencia, A., Ouzounis, C. and Sander, C. (1999a). *Bioinformatics.* **15,** 391–412.

Andrade, M. A., Ponting, C. P., Gibson, T. J. and Bork, P. (1999b). *J. Mol. Biol.* In press.

Andrade, M. A., Ouzounis, C., Sander, C., Tamames, J., and Valencia, A. (1999c). *J. Mol. Evol.* **49,** 551–557.

Aravind, L., and Koonin, E. V. (1999). *J. Mol. Evol.* **48,** 291–302.

Aravind, L., and Landsman, D. (1998). *Nucl. Acids Res.* **26,** 4413–4421.

Aravind, L., and Ponting, C. P. (1997) *Trends Biochem. Sci.* **22,** 458–459.

Aravind, L., and Ponting, C. P. (1999). *FEMS Microbiol. Lett.* **176,** 111–116.

Aravind, L., Tatusov, R. L., Wolf, Y. I., Walker, D. R., and Koonin, E. V. (1998). *Trends Genet.* **14,** 442–444.
Aravind, L., Walker, D. R., and Koonin, E. V. (1999a). *Nucl. Acids Res.* **27,** 1223–1242.
Aravind, L., Dixit, V. M., and Koonin, E. V. (1999b). *Trends Biochem. Sci.* **24,** 47–53.
Attwood, T. K., Flower, D. R., Lewis, A. P., Mabey, J. E., Morgan, S. R., Scordis, P., Selley, J. N., and Wright, W. (1999). *Nucleic Acids Res.* **27,** 220–225.
Bachhawat, A. K., Suhan, J., and Jones, E. W. (1994). *Genes Dev.* **8,** 1379–1387.
Bagby, S., Harvey, T. S., Eagle, S. G., Inouye, S., and Ikura, M. (1994). *Proc. Natl. Acad. Sci. U.S.A.* **91,** 4308–4312.
Bajt, M. L., and Loftus, J. C. (1994). *J. Biol. Chem.* **269,** 20913–20919.
Ball, L. J., Murzina, N. V., Broadhurst, R. W., Raine, A. R., Archer, S. J., Stott, F. J., Murzin, A. G., Singh, P. B., Domaille, P. J., and Laue E. D. (1997). *EMBO J.* **16,** 2473–2481.
Bargmann, C. I. (1998). Science **282,** 2028–2033.
Bateman, A., Birney, E., Durbin, R., Eddy, S. R., Finn, R. D., and Sonnhammer, E. L. (1999). *Nucleic Acids Res.* **27,** 260–262.
Baumgartner, S., Hofmann, K., Chiquet-Ehrismann, R., and Bucher, P. (1998). *Protein Sci.* **7,** 1626–1631.
Beckmann, G., Hanke, J., Bork, P., and Reich, J. G. (1998). *J. Mol. Biol.* **275,** 725–730.
Berger, K. H., and Yaffe, M. P. (1998). *Mol. Cell Biol.* **18,** 4043–4052.
Bertuch, A., and Lundblad, V. (1998). *Trends Cell Biol.* **8,** 339–342.
Birkeland, N. K. (1994). *Can. J. Microbiol.* **40,** 658–665.
Birney, E., Thompson J. D., and Gibson, T. J. (1996). *Nucl. Acids Res.* **24,** 2730–2739.
Blaxter, M. (1998). *Science* **282,** 2041–2046.
Blomberg, N., Gabdoulline, R. R., Nilges, M., and Wade, R. C. (1999). *Proteins,* **37,** 379–387.
Bork, P., and Sander, C. (1992). *FEBS Lett.* **300,** 237–240.
Bork, P., and Sudol, M. (1994). *Trends Biochem. Sci.* **19,** 531–533.
Bork, P. (1996). *Matrix Biol.* **15,** 301–310.
Bork, P., Sander, C., and Valencia, A. (1992). *Proc. Natl. Acad. Sci.* **89,** 7290–7294.
Bork, P., and Bairoch, A. (1996) *Trends Genet.* **12,** 425–427.
Bork, P., and Gibson, T. J. (1996). *Methods Enzymol.* **266,** 162–184.
Bork, P., and Koonin, E. V. (1998). *Nature Genet.* **18,** 313–318.
Bork, P., Dandekar, T., Diaz-Lascoz, Y., Eisenhaber, F., Huynen, M., and Yuan, Y. (1998). *J. Mol. Biol.* **283,** 707–725.
Borst, D. E., Redmond, T. M., Elser, J. E., Gonda, M. A., Wiggert, B., Chader, G. J., and Nickerson, J. M. (1989). *J. Biol. Chem.* **264,** 1115–1123.
Brenner, S. E. (1999). *Trends Genet.* **15,** 132–133.
Briercheck, D. M., Wood, T. C., Allison, T. J., Richardson, J. P., and Rule, G. S. (1998). *Nature Struct. Biol.* **5,** 393–399.
Brzeski, J., Podstolski, W., Olczak, K., and Jerzmanowski, A. (1999). *Nucl. Acids Res.* **27,** 2393–2399.
Bucher, P., Karplus, K., Moeri, N., and Hofmann, K. (1996). *Comput. Chem.* **20,** 3–23.
Bult, C. J., White, O., Olsen, G. J., Zhou, L., Fleischmann, R. D., Sutton, G. G., Blake, J. A., FitzGerald, L. M., Clayton, R. A., Gocayne, J. D., Kerlavage, A. R., Dougherty, B. A., Tomb, J. F., Adams, M. D., Reich, C. I., Overbeek, R., Kirkness, E. F., Weinstock, K. G., Merrick, J. M., Glodek, A., Scott, J. L., Geoghagen, N. S. M., and Venter, J. C. (1996). *Science* **273,** 1058–1073.
Butler, A. R., O'Donnell, R. W., Martin, V. J., Gooday, G. W., and Stark, M.J. (1991). *Eur. J. Biochem.* **199,** 483–488.
Butti, M. G., Bongarzone, I., Ferraresi, G., Mondellini, P., Borrello, M. G., and Pierotti, M. A. (1995). *Genomics* **28,** 15–24.

Bycroft, M., Hubbard, T. J., Proctor, M., Freund, S. M., and Murzin, A. G. (1997). *Cell* **88,** 235–242.
Callebaut, I., Courvalin, J. C., and Mornon, J. P. (1999). *FEBS Lett.* **446,** 189–193.
Castresana, J., and Saraste, M. (1995). *FEBS Lett.* **374,** 149–151.
Cavalli, G., and Paro, R. (1998). *Curr. Opin. Cell Biol.* **10,** 354–360.
Chen, J. M., Rawlings, N. D., Stevens, R. A., and Barrett, A. J. (1998). *FEBS Lett.* **441,** 361–365.
Chervitz, S. A., Aravind, L., Sherlock, G., Ball, C. A., Koonin, E. V., Dwight, S. S., Harris, M. A., Dolinski, K., Mohr, S., Smith, T., Weng, S., Cherry, J.M., and Botstein, D. (1998). *Science* **282,** 2022–2028.
Chothia, C. (1992). *Nature* **357,** 543–544.
Chung, C. Y., and Erickson, H. P. (1994). *J. Cell Biol.* **126,** 539–548.
Coates, P. J., Jamieson, D. J., Smart, K., Prescott, A. R., and Hall, P. A. (1997). *Curr. Biol.* **7,** 607–610.
Cohen, A. R., Wood, D. F., Marfatia, S. M., Walther, Z., Chishti, A. H., and Anderson, J. M. (1998). *J. Cell Biol.* **142,** 129–138.
Copley, R. R., Schultz, J., Ponting, C. P., and Bork, P. (1999). *Curr. Opin. Struct. Biol.* **9,** 408–415.
Corsaro, D., Venditti, D., Padula, M., and Valassina, M. (1999). *Crit. Rev. Microbiol.* **25,** 39–79.
Côté, J., Quinn, J., Workman, J. L., and Peterson, C. L. (1994). *Science* **265,** 53–60.
Cudmore, S., Cossart, P., Griffiths, G., and Way, M. (1995). *Nature* **378,** 636–638.
De Virgilio, C., DeMarini, D. J., and Pringle, J. R. (1996). *Microbiology* **142,** 2897–2905.
Doerks, T., Bairoch, A., and Bork, P. (1998). *Trends Genet* **14,** 248–250.
Doherty, A. J., Serpell, L. C., and Ponting, C. P. (1996). *Nucl. Acids Res.* **24,** 2488–2497.
Doolittle, R. F. (1995). *Annu. Rev. Biochem.* **64,** 287–314.
Doolittle, W. F. (1998). *Trends Genet.* **14,** 307–311.
Doolittle, W. F., and Logsdon, J. M. Jr. (1998). *Curr. Biol.* **8,** R209–R211.
Doonan, J., and Fobart, P. (1997). *Curr. Opin. Cell Biol.* **9,** 824–830.
Eschenbacher, K. H., Klein, H., Sommer, I., Meyer, H. E., Entzeroth, R., Mehlhorn, H., and Ruger, W. (1993). *Mol. Biochem. Parasitol.* **62,** 27–36.
Eudy, J. D., Weston, M. D., Yao, S-F., Hoover, D. M., Rehm, H. L., Ma-Edmonds, M., Yan, D., Ahmad, I., Cheng, J. J., Ayuso, C., Cremers, C., Davenport, S., Moller, C., Talmadge, C. B., Beisel, K. W., Tamayo, M., Morton, C. C., Swaroop, A., Kimberling, W. J., and Sumegi, J. (1998). *Science* **280,** 1753–1757.
Exposito, J. Y., and Garrone, R. (1990). *Proc. Natl. Acad. Sci. U.S.A.* **87,** 6669–6673.
Fitch, W. M. (1970). *Syst. Zool.* **19,** 99–113.
Fleischmann, R. D., Adams, M. D., White, O., Clayton, R. A., Kirkness, E. F., Kerlavage, A. R., Bult, C. J., Tomb, J. F., Dougherty, B. A., Merrick, J. M., *et al.* (1995). *Science* **269,** 496–512.
Fraser, A., and James, C. (1998). *Trends Cell Biol.* **8,** 219–221.
Frit, P., Calsou, P., Chen, D. J., and Salles, B. (1998). *J. Mol. Biol.* **284,** 963–973.
Galperin, M. Y., Walker, D. R., and Koonin, E. V. (1998). *Genome Res.* **8,** 779–790.
Gaut, B. S., and Doebley, J. F. (1997). *Proc. Natl. Acad. Sci. U.S.A.* **94,** 6809–6814.
Gertler, F. B., Niebuhr, K., Reinhard, M., Wehland, J., and Soriano. P. (1996). *Cell* **87,** 227–239.
Gibson, T. J., Thompson, J. D., and Heringa, J. (1993). *Trends Biochem. Sci.* **324,** 361–366.
Gilbert, W. (1978). *Nature* **271,** 501.
Gimona, M., and Winder, S. J. (1998). *Curr. Biol.* **24,** R674–R675.

Goffeau, A., Barrell, B. G., Bussey, H., Davis, R. W., Dujon, B., Feldmann, H., Galibert, F., Hoheisel, J. D., Jacq, C., Johnston, M., Louis, E. J., Mewes, H. W., Murakami, Y., Philippsen, P., Tettelin, H., and Oliver, S. G. (1996). *Science* **274,** 563–567.

Graumann, P. L., and Marahiel, M. A. (1998). *Trends Biochem. Sci.* **23,** 286–290.

Gray, M. W., Burger, G., and Lang, B.F. (1999). *Science* **283,** 1476–1481.

Green, P., Lipman, D., Hillier, L., Waterston, R., States, D., and Claverie, J.M. (1993). *Science* **259,** 1711–1716.

Greco, A., Mariani, C., Miranda, C., Pagliardini, S., and Pierotti, M.A. (1993). *Genomics* **18,** 397–400.

Haaning, J., Oxvig, C., Overgaard, M.T., Ebbesen, P., Kristensen, T., and Sottrup-Jensen L. (1996). *Eur. J. Biochem.* **237,** 159–163.

Han, W-D., Kawamoto, S., Hosoya, Y., Fujita, M., Sadaie, Y., Suzuki, K., Ohashi, Y., Kawamura, F., and Ochi, K. (1998). *Gene* **217,** 31–40.

Hartzell P.L. (1997). *Proc. Natl. Acad. Sci. U.S.A.* **94,** 9881–9886.

Henikoff, S., and Henikoff., J. G. (1991). *Nucl. Acids Res.* **19,** 6565–6572.

Henikoff, S., Greene, E. A., Pietrokovski, S., Bork, P., Attwood, T. K., and Hood, L. (1997). *Science* **278,** 609–614.

Henikoff, J. G., Henikoff, S., and Pietrokovski, S. (1999). *Nucleic Acids Res.* **27,** 226–228.

Hester, G., Kaku, H., Goldstein, I. J., and Wright, C. S. (1995). *Nature Struct. Biol.* **2,** 472–479.

Hirsch, J. A., Schubert, C., Gurevich, V. V., and Sigler, P. B. (1999). *Cell* **97,** 257–269.

Hofmann, K. (1998). *In* "Trends Guide to Bioinformatics", *Trends Genet. Suppl.* 18–21.

Hofmann, K., Bucher, P., Falquet, L., and Bairoch, A. (1999) *Nucleic Acids Res.* **27,** 215–219.

Hogan, L. H., Josvai, S., and Klein, B.S. (1995). *J. Biol. Chem.* **270,** 30725–30732.

Holm, L., and Sander, C. (1996). *Science* **273,** 595–602.

Holm, L., and Sander, C. (1998). *Bioinformatics* **14,** 423–429.

Hunter, T., and Plowman, G.D. (1997). *Trends Biochem. Sci.* **22,** 18–22.

Huth, J. R., Bewley, C. A., Nissen, M. S., Evans, J. N., Reeves, R., Gronenborn, A. M., and Clore, G. M. (1997). *Nat. Struct. Biol.* **4,** 657–665.

Imbalzano, A. N., Kwon, H., Green, M. R., and Kingston, R. E. (1994). *Nature* **370,** 481–485.

Inohara-Ochiai, M., Nakayama, T., Nakao, M., Fujita, T., Ueda, T., Ashikari, T., Nishino, T., and Shibano, Y. (1998). *Biochim. Biophys. Acta* **1388,** 77–83.

Janin, J., and Chothia, C. (1985). *Methods Enzymol.* **115,** 420–430.

Jansen, G., Thijssen, K. L., Werner, P., van der Horst, M., Hazendonk, E., and Plasterk, R. H. A. (1999). *Nat. Genet.* **21,** 414–419.

Jeddeloh J. A., Stokes, T. L., and Richards, E. J. (1999). *Nat. Genet.* **22,** 94–97.

Jeon, I. S., Davis, J. N., Braun, B. S., Sublett, J. E., Roussel, M. F., Denny, C. T., and Shapiro, D. N. (1995). *Oncogene* **10,** 1229–1234.

Jung, E., Fucini, P., Stewart, M., Noegel, A. A., and Schleicher, M. (1996). *EMBO J.* **15,** 1238–1246.

Kalchman, M. A., Koide, H. B., McCutcheon, K., Graham, R. K., Nichol, K., Nishiyama, K., Kazemi-Esfarjani, P., Lynn, F. C., Wellington, C., Metzler, M., Goldberg, Y. P., Kanazawa, I., Gietz, R. D., and Hayden, M. R. (1997). *Nat. Genet.* 16, 44–53.

Kanai, T., Ueki, N., Kawaguchi, T., Teranishi, Y., Atomi, H., Tomorbaatar, C., Ueda, M., and Tanaka, A. (1997). *Appl. Environ. Microbiol.* **63,** 4956–4960.

Kay, B. K., Yamabhai, M., Wendland, B., and Emr, S. D. (1999). *Protein Sci.* **8,** 435–438.

Keiler, K. C., and Sauer, R. T. (1995). *J. Biol. Chem.* **270,** 28864–28868.

Kidwell, M. G. (1993). *Annu. Rev. Genet.* **27,** 235–256.

Koonin, E. V., Mushegian, A. R., and Bork, P. (1996). *Trends Genet.* **12,** 334–336.

Koonin, E. V., Mushegian, A. R., Galperin, M. Y., and Walker, D. R. (1997). *Mol. Microbiol.* **25,** 619–637.
Koonin, E. V., Tatusov, R. L., and Galperin, M. Y. (1998). *Curr. Opin. Struct. Biol.* **8,** 355–363.
Kordis, D., and Gubensek, F. (1998). *Proc. Natl. Acad. Sci. U.S.A.* **95,** 10704–10709.
Kreikemeyer, B., Talay, S.R., and Chhatwal, G.S. (1995). *Mol. Microbiol.* **17,** 137–145.
Krogh, A., M. Brown, M., Mian, I. S., Sjolander, K., and Haussler, D. (1994). *J. Mol. Biol.* **235,** 1501–1531.
Kull, F. J., Sablin, E. P., Lau, R., Fletterick, R. J., and Vale, R. D. (1996). *Nature* **380,** 550–555.
Lee, J., Rieu, P., Arnaout, M., and Liddington, R. (1995). *Cell* **80,** 631–638.
Leonard, C. J., Aravind, L., and Koonin, E. V. (1998). *Genome Res.* **8,** 1038–1047.
Little, E., Bork, P., and Doolittle, R. F. (1994). *J. Mol. Evol.* **39,** 631–643.
Littleton, J. T., Bhat, M. A., and Bellen, H. J. (1997). *J. Cell Biol.* **137,** 793–796.
Loftus, J. C., and Liddington, R. C. (1997). *J. Clin. Invest.* **99,** 2302–2306.
Löwe, J., and Amos, L. A. (1998). *Nature* **391,** 203–206.
Lupas, A. (1997). *Curr. Opin. Struct. Biol.* **7,** 388–393.
Maclennan, A. J., and Shaw, G. (1993). *Trends Biochem. Sci.* **18,** 464–465.
Makarova, K. S., Aravind, L., Galperin, M. Y., Grishin, N. V., Tatusov, R. L., Wolf, Y. I., and Koonin, E. V. (1999). *Genome Res.* **9,** 608–628.
Marfatia, S. M., Lue, R. A., Branton, D., and Chishti, A. H. (1995). *J. Biol. Chem.* **270,** 715–719.
Margolis, L. (1970). "Origin of eukaryotic cells". Yale University Press, New Haven, CT.
May, A. P., and Ponting, C. P. (1999). *Trends Biochem. Sci.* **24,** 12–13.
McFadden, G. I., Gilson, P. R., Hofmann, C. J., Adcock, G. J., and Maier, U. G. (1994). *Proc. Natl. Acad. Sci. U.S.A.* **91,** 3690–3694.
McKnight, C. J., Matsudaira, P. T., and Kim, P. S. (1997). *Nat. Struct. Biol.* **4,** 180–184.
Mian, I. S. (1997). *Nucl. Acids Res.* **25,** 3187–3195.
Mimori, T., and Hardin, J. R. (1986). *J. Biol. Chem.* **261,** 10375–10379.
Missler, M., and Südhof, T. C. (1998). *Trends Genet.* **14,** 20–26.
Mizuno, T. (1998). *J. Biochem.* **123,** 555–563.
Moran, J. V., DeBerardinis, R.J., and Kazazian, H.H. Jr. (1999). *Science* **283,** 1530–1534.
Mukherjee A., and Lutkenhaus, J. (1994). *J. Bacteriol.* **176,** 2754–2758.
Murray, A., and Hunt, T. (1993). "The Cell Cycle," Oxford University Press, Oxford, UK.
Murzin, A. G. (1992). *Proteins* **14,** 191–201.
Murzin, A. G. (1998). *Curr. Opin. Struct. Biol.* **8,** 380–387.
Mushegian, A. R., Bassett, D. E. Jr., Boguski, M. S., Bork, P., and Koonin, E. V. (1997). *Proc. Natl. Acad. Sci. U.S.A.* **94,** 5831–5836.
Nakamura, A., Hattori, M., and Sakaki, Y. (1997). *J. Biochem.* **122,** 872–877.
Nelson, K. E., Clayton, R. A., Gill, S. R., Gwinn, M. L., Dodson, R. J., Haft, D. H., Hickey, E. K., Peterson, J. D., Nelson, W. C., Ketchum, K. A., McDonald. L., Utterback, T. R., Malek, J. A., Linher, K. D., Garrett, M. M., Stewart, A. M., Cotton, M. D., Pratt, M. S., Phillips, C. A., Richardson, D., Heidelberg, J., Sutton, G. G., Fleischmann, R. D., Eisen, J. A., White, O., Salzberg, S. L., Smith, M. O., Venter, J. C., and Fraser, C. M. (1999). *Nature* **399,** 323–329.
Neuwald, A. F., Liu, J. S., Lipman, D. J., and Lawrence, C. E. (1997). *Nucl. Acids Res.* **25,** 1655–1677.
Neuwald, A. F., Aravind, L., Spouge, J. L., and Koonin, E. V. (1999). *Genome Res.* **9,** 27–43.
Noble, J. A., Innis, M. A., Koonin, E. V., Rudd, K. E., Banuett, F., and Herskowitz, I. (1993). *Proc. Natl. Acad. Sci. U.S.A.* **90,** 10866–10870.
Nogales, E., Wolf, S. G., and Downing, K. H. (1998). *Nature* **391,** 199–203.
Ohno, S. (1970). "Evolution by Gene Duplication." Springer-Verlag, Berlin and New York.

Ornaghi, P., Ballario, P., Lena, A. M., Gonzalez, A., and Filetici, P. (1999). *J. Mol. Biol.* **287,** 1–7.

Pagano, A., Letourneur, F., Garcia-Estefania, D., Carpentier, J. L., Orci, L., and Paccaud, J. P. (1999). *J. Biol. Chem.* **274,** 7833–7840.

Park, J., Teichmann, S. A., Hubbard, T., and Chothia, C. (1997) *J. Mol. Biol.* **273,** 349–354.

Park, J., Karplus, K., Barrett, C., Hughey, R., Haussler, D., Hubbard, T., and Chothia, C. (1998). *J. Mol. Biol.* **284,** 1201–1210.

Patthy, L. (1985). *Cell* **41,** 657–663.

Patthy, L. (1996). *Matrix Biol.* **15,** 301–310.

Pearson, W. R., and Lipman, D. J. (1988). *Proc. Natl. Acad. Sci. U.S.A.* **85,** 2444–2448.

Penkett, C. J., Redfield, C., Jones, J. A., Dodd, I., Hubbard, J., Smith, R. A. G., Smith, L. J., and Dobson, C. M. (1998). *Biochemistry* **37,** 17054–17067.

Peter, M., Couturier, J., Pacquement, H., Michon, J., Thomas, G., Magdelenat, H., and Delattre, O. (1997). *Oncogene* **14,** 1159–1164.

Pohlman, R. F., Liu, F., Wang, L., More, M. I., and Winans, S. C. (1993). *Nucl. Acids Res.* **21,** 4867–4872.

Ponting, C. P. (1996). *Protein Sci.* **5,** 2353–2357.

Ponting, C. P., and Pallen, M. J. (1999). *Mol. Microbiol.* **31,** 739–740.

Ponting, C. P., Phillips, C., Davies, K. E., and Blake, D. J. (1997). *Bioessays* **19,** 469–479.

Ponting, C. P., and Parker, P. J. (1996). *Protein Sci.* **5,** 162–166.

Ponting, C. P., and Aravind, L. (1997). *Curr. Biol.* **7,** R674–R677.

Ponting, C. P., Blake, D. J., Davies, K.E., Kendrick-Jones, J., and Winder, S.J. (1996). *Trends Biochem. Sci.* **21,** 11–13.

Ponting, C. P., Schultz, J., Milpetz, F., and Bork, P. (1999a). *Nucleic Acids Res.* **27,** 229–232.

Ponting, C. P., Aravind, L., Schultz, J., Bork, P., and Koonin, E.V. (1999b). *J. Mol. Biol.* **289,** 729–746.

Ponting, C. P., Hofmann, K., and Bork, P. (1999c) *Curr. Biol.* **9,** R585–R588.

Popov, K. M., Kedishvili, N. Y., Zhao, Y., Shimomura, Y., Crabb, D. W., and Harris, R. A. (1993). *J. Biol. Chem.* **268,** 26602–26606.

Posas, F., Wurgler-Murphy, S. M., Maeda, T., Witten, E. A., Thai, T. C., and Saito, H. (1996). *Cell* **86,** 865–875.

Prehoda, K. E., Lee, D. J., and Lim, W. A. (1999). *Cell* **97,** 471–480.

Preuss, D. (1999). *Plant Cell* **11,** 765–768.

Rapraeger, A. C., and Ott, V. L. (1998). *Curr. Opin. Cell Biol.* **10,** 620–628.

Reeve, J. N., Sandman, K., and Daniels, C. J. (1997). *Cell* **89,** 999–1002.

Rivera, M. C., Jain, R., Moore, J. E., and Lake, J. A. (1998). *Proc. Natl. Acad. Sci. U.S.A.* **95,** 6239–6244.

Rost, B., and Sander, C. (1993). *J. Mol. Biol.* **232,** 584–599.

Russell, R. B. (1998). *J. Mol. Biol.* **279,** 1211–1227.

Ruvkun, G., and Hobert, O. (1998). *Science* **282,** 2033–2041.

Salamov, A. A., Suwa, M., Orengo, C. A., and Swindells, M. B. (1999). *Protein Eng.* **12,** 95–100.

Satterlee, J. S., and Sussman, M. R. (1998). *J. Membr. Biol.* **164,** 205–213.

Scheffzek, K., Ahmadian, M. R., and Wittinghofer, A. (1998) *Trends Biochem. Sci.* **23,** 257–262.

Schieven, G., Thorner, J., and Martin, G.S. (1986). *Science* **231,** 390–393.

Schindelin, H., Jiang, W., Inouye, M., and Heinemann, U. (1994). *Proc. Natl. Acad. Sci.* **91,** 5119–5123.

Schleicher, M., Andre, E., Hartmann, H., and Noegel, A. A. (1988). *Dev. Genet.* **9,** 521–530.

Schultz, J., Milpetz, F., Bork, P., and Ponting, C. P. (1998) *Proc. Natl. Acad. Sci. U.S.A.* **95**, 5857–5864.
Schwarz, E. M., and Benzer, S. (1997). *Proc. Natl. Acad. Sci. U.S.A.* **94,** 10249–10254.
Sebastiano, M., Lassandro, F., and Lazzicalupo, P. (1991). *Dev. Biol.* **146,** 519–530.
Shapiro, L., and Lima, C. D. (1998). *Structure* **6,** 265–267.
Shaw, G. (1996). *Bioessays* **18,** 35–46.
Sidow, A. (1996). *Curr. Opin. Genet. Dev.* **6,** 716–722.
Silber, K. R., Keiler, K. C., and Sauer, R.T. (1992). *Proc. Natl. Acad. Sci. U.S.A.* **89,** 295–299.
Slack, F. J., and Ruvkun, G. (1998). *Trends Biochem. Sci.* **23,** 474–475.
Sluder, A. E., Mathews, S. W., Hough, D., Yin, V. P., and Maina C. V. (1999). *Genome Res.* **9,** 103–120.
Smith, T. F., and Waterman, M. S. (1981). *J. Mol. Biol.* **147,** 195–197.
Smith, T. F., and Zhang, X. (1997). *Nat. Biotechnol.* **15,** 1222–1223.
Snow, B. E., Krumins, A. M., Brothers, G. M., Lee, S. F., Wall, M. A., Chung, S., Mangion, J., Arya, S., Gilman, A. G., and Siderovski, D. P. (1998). *Proc. Natl. Acad. Sci. U.S.A.* **95,** 13307–13312.
Solomon, M. J. (1993). *Curr. Opin. Cell Biol.* **5,** 180–186.
Spraggon, G., Everse, S. J., and Doolittle, R. F. (1997). *Nature* **389,** 455–462.
Steglich, G., Neupert, W., and Langer, T. (1999). *Mol. Cell. Biol.* **19,** 3435–3442.
Stephens, R. S., Kalman, S., Lammel, C., Fan, J., Marathe, R., Aravind, L., Mitchell, W., Olinger, L., Tatusov, R. L., Zhao, Q., Koonin, E. V., and Davis, R. W. (1998). *Science* **282,** 754–759.
Sun, Z., Hsiao, J., Fay, D. S., and Stern, D. F. (1998). *Science* **281,** 272–274.
Swan, D. G., Cortes, J., Hale, R. S., and Leadlay, P. F. (1989). *J. Bacteriol.* **171,** 5614–5619.
Tamura, T., Tamura, N., Cejka, Z., Hegerl, R., Lottspeich, F., and Baumeister, W. (1996). *Science* **274,** 1385–1389.
Tatusov, R. L., Altschul, S. F., and Koonin, E. V. (1994). *Proc. Natl. Acad. Sci. U.S.A.* **91,** 12091–12095.
Tatusov, R. L., Koonin, E. V., and Lipman, D. J. (1997). *Science* **278,** 631–637.
Taylor, B. L., and Zhulin, I. B. (1999). *Microbiol. Mol. Biol. Rev.* **63,** 479–506.
Thompson, J. D., Higgins, D. G., and Gibson, T. J. (1994). *Nucl. Acids Res.* **22,** 4673–4680.
Tozer, E. C., Liddington, R. C., Sutcliffe, M., Smeeton, A. H., and Loftus, J. C. (1996). *J. Biol. Chem.* **271,** 21978–21984.
Tuckwell, D. S., and Humphries, M. J. (1997). *FEBS Lett.* **400,** 297–303.
Udo, H., Munoz-Dorado, J. Inouye, M., and Inouye, S. (1995). *Genes Dev.* **9,** 972–983.
Van Troys, M., Dewitte, D., Goethals, M., Carlier, M-F., Vandekerckhove, J., and Ampe, C. (1996). *EMBO J.* **15,** 201–210.
Van Troys, Vanderkerckhove, J., and Ampe, C. (1999). *Biochim. Biophys. Acta* **1448,** 323–348.
Verna, J., Lodder, A., Lee, K., Vagts, A., and Ballester, R. (1997). *Proc. Natl. Acad. Sci. U.S.A.* **94,** 13804–13809.
Wang, J., Dong, X., Myung, K., Hendrickson, E.A., and Reeves, W.H. (1998). *J. Biol. Chem.* **273,** 842–848.
Wanker, E. E., Rovira, C., Scherzinger, E., Hasenbank, R., Walter, S., Tait, D., Colicelli, J., and Lehrach, H. (1997). *Hum. Mol. Genet.* **6,** 487–495.
Wen, D., Corina, K., Chow, E. P., Miller, S., Janmey, P. A., and Pepinsky, R.B. (1996). *Biochemistry* **35,** 9700–9709.
Wesp, A., Hicke, L., Palecek, J., Lombardi, R., Aust, T., Munn, A. L., and Riezman, H. (1997). *Mol. Biol. Cell* **8,** 2291–2306.
Whisstock, J. C., and Lesk, A. M. (1999). *Trends Biochem. Sci.* **24,** 132–133.

Widmann, C., Gibson, S., Jarpe, M. B., and Johnson, G. L. (1999). *Physiol. Rev.* **79,** 143–180.
Wilson, C., Goberdhan, D. C., and Steller, H. (1993). *Proc. Natl. Acad. Sci. U.S.A.* **90,** 7109–7113.
Woese, C. (1998). *Proc. Natl. Acad. Sci. U.S.A.* **95,** 6854–6959.
Wolf, Y. I., Aravind, L., and Koonin, E. V. (1999a). *Trends Genet.* **15,** 173–175.
Wolf, Y., Brenner, S. E., Bash, P. A., and Koonin, E. V. (1999b). *Genome Res.* **9,** 17–26.
Wolfe, K. H., and Shields, D. C. (1997). *Nature* **387,** 708–713.
Wu, A. L., Hallstrom, T. C., and Moye-Rowley, W. S. (1996). *J. Biol. Chem.* **271,** 2914–2920.
Wu, Y-C., and Horvitz, H. R. (1998). *Nature* **392,** 501–504.
Wu, D. M., Zhang, Y., Parada, N. A., Kornfeld, H., Nicoll, J., Center, D. M., and Cruikshank, W. W. (1999). *J. Immunol.* **163,** 1287–1293.
Xu, X., and Doolittle, R. F. (1990). *Proc. Natl. Acad. Sci. U.S.A.* **87,** 2097–2101.
Xu, X. Z., Wes, P. D., Chen, H., Li, H. S., Yu, M., Morgan, S., Liu, Y., and Montell, C. (1998). *J. Biol. Chem.* **273,** 31297–31307.
Yamada, T., Fukuda, R., Himeno, M., and Sugimoto, K. (1999). *J. Biochem.* **125,** 832–837.
Yang, D., Oyaizu, Y., Oyaizu, H., Olsen, G. J., and Woese, C. R. (1985). *Proc. Natl. Acad. Sci. U.S.A.* **82,** 4443–4447.
Zavitz, K. H., and Zipursky, S. L. (1997). *Curr. Opin Cell Biol.* **9,** 773–781.
Zhulin, I. B., Taylor, B. L., and Dixon, R. (1997). *Trends Biochem. Sci.* **22,** 331–333.
Zimmermann, P., and David, G. (1999). *FASEB J.* **13** (Suppl)., S91–S100.