SEQUENCE ANALYSIS OF MULTIDOMAIN PROTEINS: PAST PERSPECTIVES AND FUTURE DIRECTIONS

BY RICHARD R. COPLEY,* CHRIS P. PONTING,† JÖRG SCHULTZ,*,‡,[1] AND PEER BORK*,‡

* EMBL, 69012 Heidelberg, Germany, †MRC Functional Genetics Unit, Department of Human Anatomy and Genetics, University of Oxford, Oxford OX1 3QX UK, United Kingdom, and ‡Max-Delbrück-Center for Molecular Medicine, Berlin-Buch, Germany

I. Identification of Novel Protein Domain Families

The reductionist concept of domains in proteins has important roles to play in structural biology, genetics, evolution, and biochemistry. Unfortunately, however, this concept often is used differently in each of these subdisciplines. In structural biology, protein domains are usually defined as continuous polypeptide chains that are folded into spatially distinct structural units (e.g., Janin and Chothia, 1985). By contrast, a domain is often defined in the biochemical and genetic literature as the minimal fragment of a gene that is still able to perform a certain function, such as that identified in deletion experiments. In sequence analysis, domains are usually defined as such only when they are contiguous in sequence and when they are found in different multidomain contexts, for example, when they occur with different flanking domains.

The situation in which domain homologs are present in proteins with different domain compositions and arrangements is assumed to have arisen through intragenomic duplication and recombination events. Such genetically mobile domains are also sometimes called modules. This term was originally introduced into the protein world in the context of immunoglobulin domains, but was later used to describe packed

[1]Present address: CellZome, Meyerhofstrasse 1, 69117 Heidelberg, Germany.

ADVANCES IN PROTEIN CHEMISTRY, Vol. 61

0065-3233/03 $35.00

supersecondary structure elements within the context of exon-shuffling theories of gene evolution (Go, 1983). Later still, however, the term was used to describe mobile domains in extracellular multidomain proteins (Patthy *et al.*, 1984). Use of this term has increased rapidly since the beginning of the 1990s (e.g., Baron *et al.*, 1991; Bork, 1992). Modules are viewed as evolutionarily independent entities that may be found in single copies. As such they differ from repeats, sequence units that are structurally and functionally interdependent and require multiple copies to form a stable structure which, in turn, may be considered as a domain. A third commonly used term is "motif." This may encompass only a portion of a domain, such as active or binding site residues, or may exist outside of domains in sequence regions that are structured only when bound to a substrate. For further definitions of terms, see reviews such as Bork and Koonin (1996).

Analysis of completed genomes shows that eukaryotes encode larger numbers of longer proteins (Fig. 1) As structural domains typically have an average size of approximately 100–150 residues, this leads to the conclusion that a greater proportion of proteins found in eukaryotes are multidomain in character when compared to prokaryotes. As each domain is likely to contribute differently to the functional attributes of the protein as a whole, it follows that identification of domains, repeats, and motifs is often an essential part of understanding how higher levels of protein function emerge during the processes of evolution.

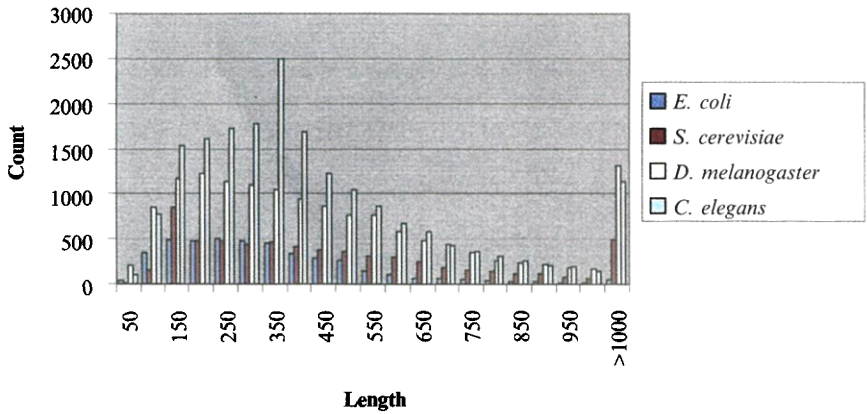

FIG. 1. Histogram of the lengths of predicted proteins encoded in ths completely sequenced genomes of *Esherichia coli*, *Saccharomyces cerevisiae*, *Caenorhabditis elegans*, and *Drosophila melanogaster*.

Functional attributes of a protein, such as active or binding sites, can be assigned to domains, repeats, and motifs. Other structures, however, also contribute to a protein's functions. These are sequences that target proteins to, for example, membranes (transmembrane helices) or organelles (such as signal peptides or mitochondrial import sequences). Coiled coils can contribute to function by mediating protein–protein interactions or spatially separating functionally distinct protein regions. By understanding the functions of the individual regions of a sequence, such as domains, motifs, coils, transmembrane helices, signal peptides, and other targeting sequences, we aim to move toward an understanding of the protein as a whole.

Here we review different strategies of domain identification at the sequence level from a historical perspective, point to some future directions of domain research, and describe domain discovery in the context of genome analysis. To support our points, we provide illustrative examples of domains that are mostly represented in SMART (Simple Modular Architecture Research Tool; Schultz *et al.*, 1998, 2000). Where a SMART domain name is mentioned in the text, we represent it in boldface.

A. History of Domain Discovery

Following determination of the first crystal structures of proteins, it became obvious that large proteins, such as dehydrogenases, are frequently composed of different structural units or domains (Adams *et al.*, 1970). Although often containing different enzymatic domains, NAD(P)-dependent dehydrogenases were found to possess a common dinucleotide-binding domain (Rossmann *et al.* 1974). From these and similar studies it soon became apparent that pairs of domains do not necessarily always co-occur in proteins.

With greater numbers of sequences and structures becoming known in the 1970s, investigators became increasingly reliant on sequence to predict structure. At this time it became a matter of faith that pairs of sequences with considerable similarity possess highly similar structures. Sequence comparisons thus acquired greater significance, and algorithms were devised for alignment (Needleman and Wunsch, 1970; Smith and Waterman, 1981). This gave rise to the thorny question of whether observed sequence similarities imply an evolutionary relatedness or else a chance event (see the review by Altschul and Gish, 1996). This question prompted still ongoing advances in database searching algorithms (Karlin and Altschul, 1990; Altschul *et al.*, 1997; Pearson, 1998; Mott, 2000). These algorithms, in particular the BLAST suite (Altschul *et al.*, 1997), provide reliable and robust alignment score statistics that have held the key to the detection of remote homologies.

With the observation of domains in distinct molecular contexts came proposals for genetic mechanisms for their spread. Gilbert (1978, 1985) proposed that individual exons code for domains and that genes represent collections of exons brought together by recombination within intron sequences. For the specific case of extracellular proteins, this proposal gained support by constraints on the intron/exon boundaries (Patthy, 1987). There is, however, little evidence that exons code for domains in intracellular proteins (Bork, 1996).

The emphasis on exon shuffling in the late 1970s and 1980s was due, in part, to the considerable numbers of extracellular domain families that were first identified in these years (Doolittle, 1985; Patthy, 1985). Extracellular domain families have continued to be discovered, although at a greatly reduced rate in more recent years [for example, the PAN domain (Tordai *et al.*, 1999)]. To date, there are more than 150 distinct modules, occurring in extracellular portions of proteins, that have been catalogued (see, e.g., the SMART resource; Schultz *et al.*, 2000), and proposals for their nomenclature have been made (Bork and Bairoch, 1995).

By contrast, the majority of intracellular domains were only identified in the 1990s (Bork *et al.*, 1997) (Fig. 2a). These domains are mostly involved in signaling, transport, and nuclear processes, but they can also flank the catalytic domains of metabolic enzymes. Identification of intracellular signaling domain families, which are frequently considerably diverged in sequence, was greatly facilitated by improvements in database search algorithms such as BLAST. When one considers, however, the number of proteins currently known to contain these domains (Fig. 2b), it appears that most of the truly widespread cytoplasmic signaling domain families have already been discovered. This is not meant to imply that the era of important domain discoveries is over, only that it may be more profitable in the future to consider the domains of cellular processes other than cytoplasmic signaling.

The above discussion makes a simplistic distinction between extracellular and intracellular domains. Although many domain families such as kringle (**KR**), epidermal growth factorlike (**EGF**), and fibronectin type I and II (**FN1**, **FN2**) domains appear to occur *only* in extracellular proteins, for other domains this is not always the case. Indeed, fibronectin type III (**FN3**), von Willebrand factor A (**VWA**), and immunoglobulin domains (**IG**), often described as "extracellular domains," are known in intracellular proteins. A more extreme example is that of **PDZ** domains, most frequently seen in intracellular proteins, but which have also been identified in nuclear and in extracellular proteins (SIP-1 and interleukin-16, respectively).

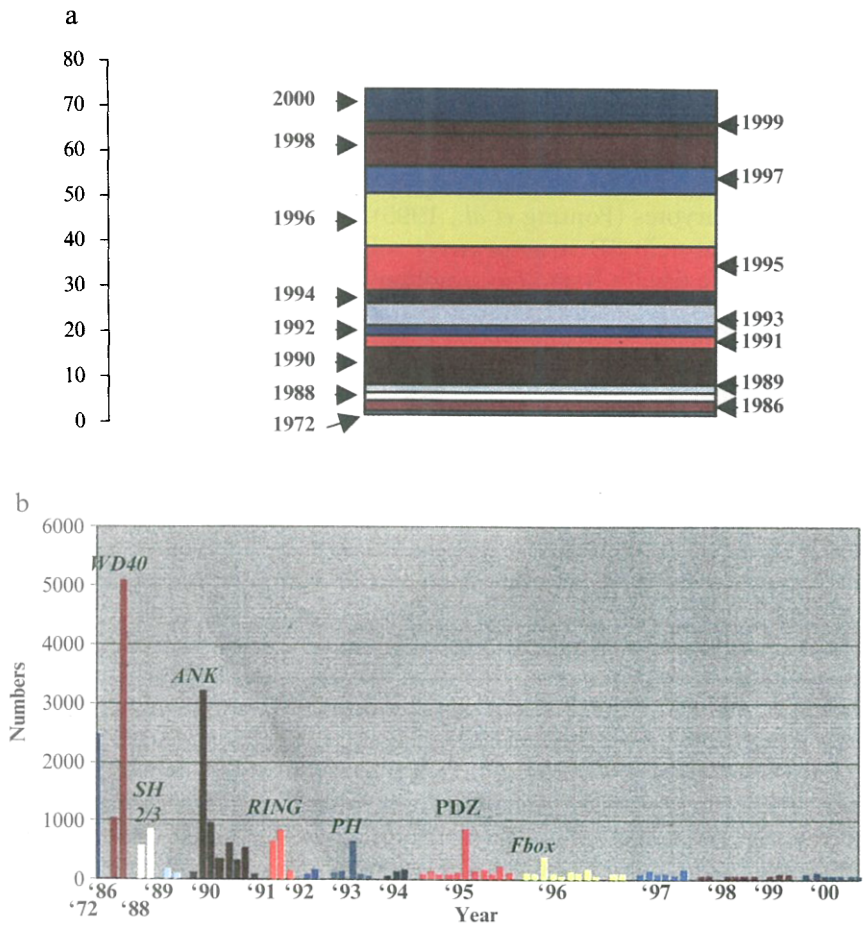

FIG. 2. (a) Proportions of currently known signaling domains discovered in a particular year. (b) Chronology of discovery of cytoplasmic signaling domains. Each year is colored distinctly. The heights of the bars represent the number of proteins that the domain is found in. Particularly common domains are labeled.

B. Domain Discovery Today

The rapid expansion of sequence and structure databases is proving to be a great boon to domain discovery and our understanding of domain propagation. Knowledge of complete genome sequences from several representatives of all three forms of cellular life (archaea, bacteria, and eukarya) has focused attention on the presence of prokaryotic homologs

of domains previously thought to be specific to eukaryotes. From current phylogenetic distributions, it has been possible to infer the presence of the majority of enzymatic domains in the last common ancestor ("cenancestor") of archaea, bacteria, and eukarya (Ponting *et al.*, 1999). Equally, attempts to identify chordate-specific enzymes have produced only a single example thus far (Lander *et al.*, 2001). In contrast, the majority of signaling domains in eukaryotes appear to have no homologs in the prokaryotes (Ponting *et al.*, 1999). Such conclusions may require revision as protein 3D structures are solved, and it becomes possible to identify more distant homologs. Another theme that has recently become apparent is the relatively frequent horizontal transfer of genes encoding signaling domains from eukaryotes to bacteria (but not to archaea) (Ponting *et al.*, 1999). One pressing issue that has arisen from these studies and remains to be resolved is whether such genes have acquired functions in the new bacterial contexts that are distinct from their eukaryotic counterparts.

The sequence and structure data explosion is prompting a greater realization that domain families, once thought to be evolutionarily distinct, might merely be separate branches of a greater superfamily tree. Examples of this abound. Extracellular families of tumor necrosis factor (**TNF**) and complement 1q (**C1Q**) are now thought to be homologous (Shapiro and Scherer, 1998), as are the intracellular families of WASp-homology 1 (**WH1**) and Ran-binding domains (**RanBD**) (Callebaut *et al.*, 1998). Even functionally distinct molecules, such as the extracellular cytokines, interleukins-1, and fibroblast growth factors and the intracellular actin-binding proteins, hisactophilin and fascin, have been shown to be distant homologs (Ponting and Russell, 2000). The "merging" of sequence families into superfamilies is also being increasingly seen among the repeats. For example, HAT, protein farnesyl transferase A, and SEL-1 repeats are all now recognized as divergent subfamilies of tetratricopeptide repeats (**TPR**) (Andrade *et al.*, 2000; Ponting, 2000).

Assignment of distant homology is now being greatly assisted by the large numbers of three-dimensional structures of modules being determined, as proteins can share similar structures even though sequence similarity is undetectable with current methods. Several NMR and crystallography groups are specializing in solving structures of modules that are classified in domain databases (e.g., Tsujishita and Hurley, 2000; see below). This has led to the current situation where 65% of the domain families in the SMART database (Schultz *et al.*, 2000) have at least one family member whose structure has been determined and deposited in the PDB.

II. Methods for Classifying Protein Domain Families

The process of collating homologous members of a domain family can be achieved using either automatic clustering "top-down" methods or else a "bottom-up" method of combining automatic searches with expert assessment of results. The automatic clustering methods build multiple alignments on the basis of "all-against-all" sequence comparisons of a given database (or genome). This approach is much faster and more systematic than bottom-up semiautomatic methods but has significant limitations in sensitivity and selectivity since, although many approaches have been tried (Heger and Holm, 2000), they often provide inaccurate domain boundaries and alignments. For a description of a frequently used top-down approach, the reader is referred to a recent report of improvements to the ProDom database (Corpet *et al.*, 2000).

Bottom-up semiautomatic approaches curate individual families by establishing significant similarities among a set of sequence regions. These regions are used to build multiple sequence alignments with due attention paid to domain boundaries and other aspects of alignment accuracy, such as gap placement and consistency with predicted secondary structure. Such high-quality sequence alignments are valuable in themselves, as with suitable algorithmic processing they can be used to sensitively search sequence databases and identify further more distantly related homologs. Domain families are then annotated with respect to known structure and function. These approaches can improve on automatic methods if problems associated with sequence errors arising from, for example, misassembly of genes from genomic sequences or sequence fragments are resolved by expert hand-curation. On the other hand, the top-down methods have the advantage of clustering a far greater proportion of sequence databases than semiautomatic methods. The two approaches need not be mutually exclusive. By using a top-down approach on sequences with domains defined by a bottom-up approach, it is possible to rapidly screen for novel protein domains (Doerks *et al.*, 2002).

In databases built using bottom-up approaches, any computational representation believed to be common to all members of a particular domain family can be used. This representation, in conjunction with appropriate searching software, should optimally be able to distinguish all true family members from the background noise of unrelated proteins stored in sequence databases. This is a challenging problem, tackled with varying degrees of sophistication by different approaches. At the most basic level, the representation can consist of a simple pattern of amino acids common to a particular domain. Such an approach is found in

many of the PROSITE database motifs (Hofmann *et al.*, 1999). At the other extreme of sophistication are hidden Markov models (HMMs) and generalized profiles. The latter are position-specific scoring tables calculated from multiple sequence alignments (Gribskov *et al.*, 1987) and are formally equivalent to certain types of HMMs (Bucher *et al.*, 1996). HMM profiles have a strong theoretical basis for their treatment of gaps (insertion/deletion positions) and for providing robust estimators of the biological significance of sequence similarities. Consequently, these are the representations of choice for the SMART (http://smart.embl-heidelberg.de/), PFAM (http://www.sanger.ac.uk/Pfam/), and Prosite profile (http://www.isrec.isb-sib.ch/software/PFSCAN_form.html) databases.

SMART and PFAM also provide additional predictions on their servers that are unrelated to domains. These resources use coiled-coil, signal peptide, transmembrane helix, and compositional bias prediction algorithms to provide added value to their service. Other such "meta-sites" will undoubtedly flourish in the coming years. Already, the *InterPro* project (http://ebi.ac.uk/interpro) has integrated different domain databases and their respective search methods into a unified tool for proteome annotation. Currently, the underlying databases represented are PFAM, PRINTS, PROSITE, ProDom, and SMART, although other databases are likely to be integrated at a later date. The derived *InterPro* database has already been used to annotate the genome sequences of *Drosophila melanogaster* (Adams *et al.*, 2000) and *Homo sapiens* (Lander *et al.*, 2001). A major advantage of *InterPro* is that it has allowed the pooling of labor-intensive annotation efforts. Thus, literature and biochemical data on a particular domain family can be easily shared irrespective of the computational techniques used for their identification.

Another recently provided meta-site is that of CDD (Conserved Domain Database and search service) provided by the National Center for Biotechnology Information (http://web.ncbi.nlm.nih.gov/Structure/cdd/cdd.shtml). CDD uses SMART- and PFAM-derived multiple alignments, as well as additional alignments provided locally, to generate position-specific score matrices. These matrices may be compared against single sequences using a derivative of the popular BLAST suite of programs. The strengths of this approach are threefold: (i) the search algorithm is familiar to users of BLAST; (ii) results are comparable with known tertiary structures using Cn3D; and (iii) results partially complement those produced by PFAM and SMART since a different search algorithm is employed. On the other hand, the HMM-based methodology of PFAM and SMART significantly outperforms that of BLAST for short repeats or domains, and the CDD alignment search set is a subset of the union of the PFAM and SMART set. Consequently, it is recommended

that users of these resources search each of them to obtain maximum benefit.

Finally, it is important to emphasize that none of these resources will result in the detection of *all* homologs and none is guaranteed to discriminate against all nonhomologs. This is because there is always a small but finite chance that a nonhomologous sequence will be detected with a relatively high alignment score or, conversely, that a homologous sequence will be detected with a low score. A healthy degree of skepticism, therefore, is appropriate in interpreting the predictions of these resources.

Domains are frequently categorized in terms of their structure and evolution, but rarely by function. The structural classification resources such as SCOP (Lo Conte *et al.*, 2000) and CATH (Pearl *et al.*, 2000) are invaluable in deciding which domains possess the same fold and hence, implicitly, which are likely to be homologous. Similarly, homology classification resources, such as SMART, PFAM, and ProDom, predict evolutionary relationships explicitly and structural similarities implicitly (Fig. 3). Although merging domain families into superfamilies is vital to an understanding of evolution, it is important to realize that it says nothing about function explicitly. Thus the broad evolutionary approach of looking for distant homologs must be complemented by attempts to make more precise functional predictions, if such resources are to confer the greatest benefit to biology. Of course, for close evolutionary relationships, it is reasonable to assume some kind of functional relationship (Wilson *et al.*, 2000).

For a family of functionally diverse homologs, predicting function requires the use of more than one domain representation, such as a multiple alignment. Instead, one or both of two stratagems may be employed. First, a multiple alignment might be constructed that represents all branches of the superfamily tree. Partitioning into functional groups might be achieved by considering the conservation, or otherwise, of patterns of important amino acids that act as functional determinants (see below). Second, several multiple alignments that may each represent a major branch of the superfamily tree might be constructed. Function prediction relies on a homolog being more similar to sequence members of one alignment than it is to others from other alignments.

These are complementary stratagems and one is not necessarily preferable to another. On the one hand, the pattern-based approach is useful only if the functional determinants for families have been experimentally derived. On the other hand, the multiple alignments approach is useful only if it is assumed that the most sequence-similar proteins possess the most similar functions. In this regard, it is emphasized that a difference of only a single residue, for example, that in an active site, between two

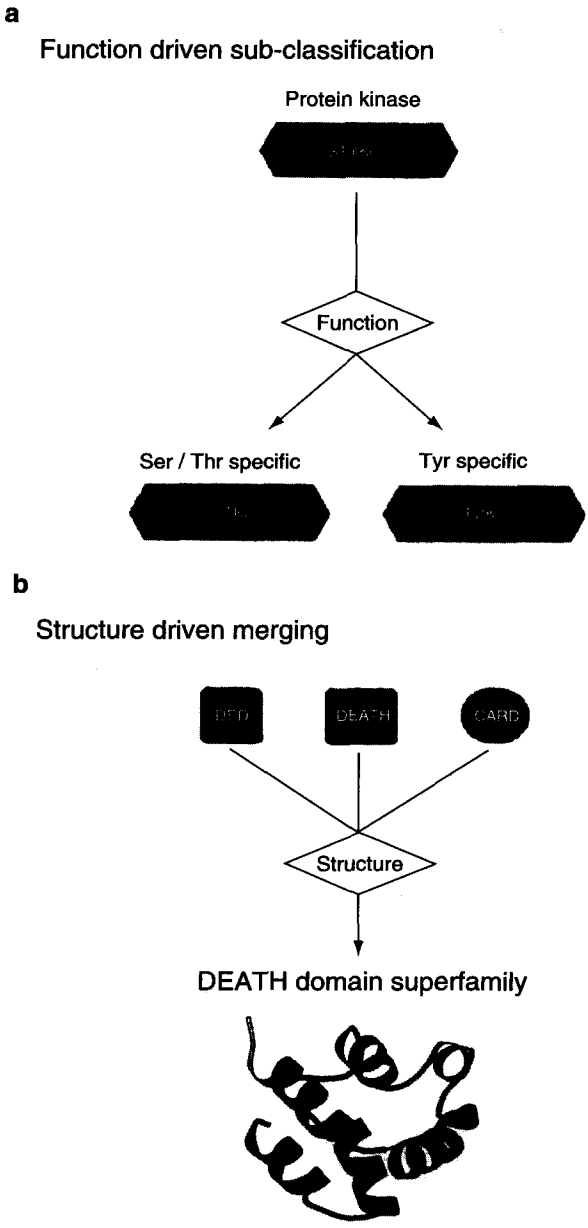

FIG. 3. (a) Divergent evolution can give rise to scenarious in which homologous domains can have different functions. If different functions are known initially, profiles or HMMs can be created that reflect the specific subfunctions of family members. SMART contains both Ser/Thr- and Tyr-specific kinase HMMs. If a sequence is a strong

sequences might dramatically alter function, even if the rest of the proteins' sequences are absolutely conserved. A striking example of this is the serine proteases of the venom of Crotalinae snakes (Deshimaru *et al.*, 1996). These possess enzyme specificities that are akin to a broad range of mammalian serine proteases. Their sequences, however, are all similar to only a small subset of these serine proteases. It is likely that considerable selective pressures on these snakes have driven an accelerated evolution of these proteases' sequences. Neither is the converse case necessarily straightforward. If two proteins share similar functions and are known to be homologous, it does not necessarily mean that they have recently diverged. For instance, the Zn^{2+}-peptidase superfamily shows independent evolution of N-deacylation and N-desuccinylation within separate lineages, pointing to functional convergence of homologous proteins (Makarova and Grishin, 1999).

Automated prediction resources, such as SMART, currently rely on multiple alignments for predicting functions of superfamily members. This is most apparent for the superfamilies of Ras-like small GTPases and protein kinases. In the future, as annotation resources mature, it is likely that combinations of cross-linked multiple alignments and patterns will be used to partition superfamilies into functionally distinct sets. This will be applicable for almost all domains, the exceptions being domain homologs that are circularly permuted or that are inserted (Russell and Ponting, 1998) and others that contain small regions of significant sequence similarity embedded in different nonhomologous contexts (Lupas *et al.*, 2001).

III. From Domain Classification to Domain Context

In the absence of experimental information, useful clues to predict the functions of the constituent domains of a protein can be acquired from several different sources. First, however, it is instructive to distinguish between functions acting at different linear scales. Predicting *domain* function requires consideration of a domain's multiple alignment at the residue scale. Predicting *protein* function, on the other hand, may involve a synthesis of its domains' functions. Conversely, some insight into the functions of a domain family may be gleaned from consideration of co-

---

enough match to either of these, it receives a specific prediction; otherwise, it is classified into the more generic **STYKc** class, corresponding to protein kinases with specificity unassigned. (b) An alternative scenario. For the **DED**, **DEATH**, and **CARD** domains, homology is not apparent from sequence alone. Only with the determination of three-dimensional structure is it possible to assign them as members of the same superfamily.

occurring domains within the context of the known function of a protein. Each of these directions is worth pursuing and undoubtedly will bring a greater functional insight into current domain collections in the near future.

A. Zooming In: Residue Context and Functional Subtyping

Since the 1980s it has been standard practice to assemble multiple alignments of superfamilies from as many family members as possible. This increases sequence variability and enables the identification of conserved functional residues. From the 1990s, several small-scale attempts have been made to identify residues characteristic of families that all share a subset of the functions of the superfamily. It has been, however, only from the mid-1990s that automatic and large-scale methods have been developed for functional subtyping (e.g., Casari *et al.*, 1995; Lichtarge *et al.*, 1996; Sjolander, 1998), and further refinements of such approaches continue to be published (Hannenhalli and Russell, 2000). Such methods typically attempt to correlate specific amino acids in a sequence alignment with particular groups of sequences (for instance, with the sequence groupings found in a phylogenetic tree).

In the absence of structural information, these methods work best for enzyme families with conserved active sites and a number of known experimental constraints on different substrates. By contrast, most mobile regulatory domain families possess subtle binding determinants that are unable to be discriminated using sequence information alone. Integration of structure with sequence information, however, has led to functional subtyping for a few regulatory domain families, such as **RA** (Kalhammer *et al.*, 1997), **WW** (Espanel and Sudol, 1999), src homology 2 (**SH2**) (Kimber *et al.*, 2000), and pleckstrin homology (**PH**) (Isakoff *et al.*, 1998) domain families.

Three-dimensional structures, if available, provide an even greater potential for identifying functionally related subfamilies from among superfamily homologs. Structures allow predictions using features that cannot be quantified from a multiple alignment, such as electrostatic potentials. This approach was used, for example, to identify a subfamily of classical **PH** domains unlikely to share the lipid-binding function ascribed to some **PH** domains (Blomberg and Nilges, 1997).

B. Zooming Out: Domain Context within Proteins

Valuable functional information may be extracted from the composition and order of domains in proteins. This arises from a basic premise that domains involved in similar cellular functions are more likely to be

found together in a multidomain protein than are functionally dissimilar domains. An example of this is the detection of **PH** domains at the C-termini of 82% of guanine nucleotide exchange factors (GEFs) acting on Rho-type small GTPases. It is predicted that all such **RhoGEFs** possess C-terminal **PH** domains, but that for the 18% minority, they are not detectable computationally, as they lack statistically significant sequence similarity. As suggested from the crystal structure of the **RhoGEF/PH** domain pair (Soisson *et al.*, 1998), the functions of these two domains are likely to be highly cooperative.

At the other extreme, some domains are found to be anti-correlated. This demonstrates that some domain types occur only in proteins found in specific cellular compartments and supports labels such as "secreted" or "nuclear" to be assigned to domain families. For example, the secreted domain **KR** is never found in tandem with a "cytoplasmic" **PH** domain. More surprisingly, some domains found in proteins targeted to the *same* cellular compartment can be anti-correlated. The most striking example of this relates to **SH2** and **PDZ** (PSD-95, Dlg, ZO-1/2) domains that *never* co-occur in sequenced proteins. This is despite a relatively high co-occurrence of both domain types with **SH3** domains: 24 and 9% of **SH3** domain-containing proteins contain **SH2** domains or **PDZ** domains, respectively. The reason for this striking negative correlation remains unknown, particularly as 16 examples are known of **PDZ** domains co-occurring with **PTB** (phosphotyrosine-binding) domains that, in some cases, have been shown to possess a function similar to that of **SH2** domains.

Other negative correlations are simpler to interpret. Two functionally antagonistic enzymes, namely, protein kinases and protein phosphatases, have, to date, not been found in the same protein. Similarly, **WW** and **SH3** domains that both bind the similar polyproline-containing substrates are never found together. This last finding, however, is curious since 216 proteins that contain either two or more **WW**, or two or more **SH3**, domains are known. Finally, it appears that proteins with domains that bind phosphoserine or phosphothreonine (**FHA**, forkhead-associated domains) never contain domains that bind phosphotyrosine (**SH2**, **PTB**, and **PTBI** domains). This indicates that cytoplasmic signaling via phosphoserine or phosphothreonine occurs via pathways distinct from those signaling via phosphotyrosine.

In 1999, the co-occurrence of domains was used to hint at the cellular functions of proteins (Marcotte *et al.*, 1999a,b; Enright *et al.*, 1999); an example of this is given in Fig. 4. These represented some of the first instances where function was inferred without explicit use of homology. Such approaches result in considerable error rates due to protein modules that can be fused with many other domains. Consequently,

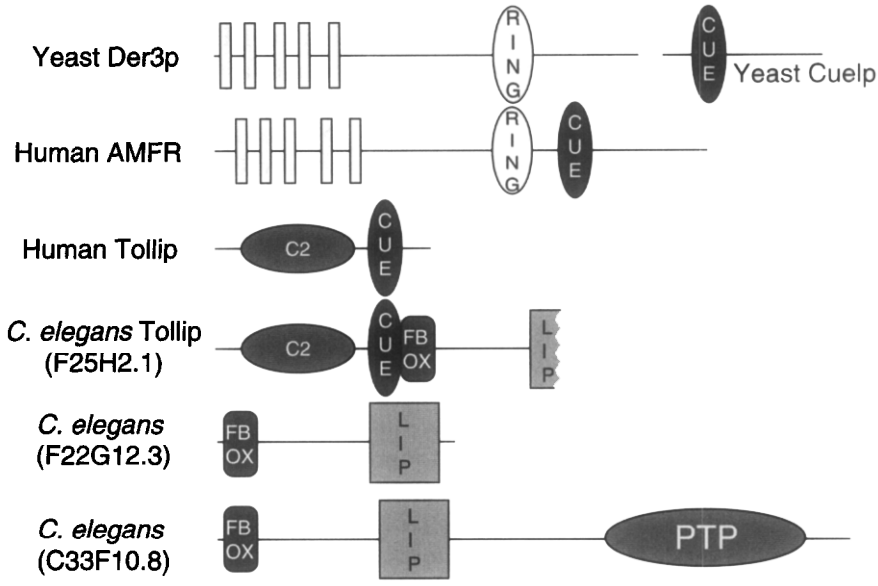

FIG. 4. The domain organizations of some CUE and LIP domain-containing proteins. Yeast Der3p/Hrd1p and Cue1p are proteins of the endoplasmic reticulum degradation pathway. As human autocrine motility factor receptor (AMFR) contains the same domain organization of a conceptual Der3p/Hrd1p and Cue1p fusion, it is proposed that Der3p/Hrd1p and Cue1p interact physically (Ponting, 2000). The *C. elegans* sequence most similar to human Tollip contains a C-terminal extension containing an F-box domain and an incomplete LIP domain. Over 190 LIP domains occur in at least 172 *C. elegans* hypothetical proteins, but have not been observed in other species' sequences; the functions of this domain remain unknown. LIP domains frequently co-occur with F-box domains and in one case (C33F10.8) a protein tyrosine phosphatase-like (PTP) domain.

these "promiscuous" domains were discarded in the prediction procedure (Marcotte *et al.*, 1999a). Even when such fusions occur, there may be little that can be usefully said about function. Before the current round of systematic analyses Pekarsky *et al.* (1998) recognized the significance of *Caenorhabditis elegans* and *Drosophila* fusion proteins (since termed "Rosetta" sequences) of Fhit, a human tumor suppressor gene of unknown function, and Nit, a member of a protein family homologous to nitrilases. As little was known about this latter family, little light was shed on the function of Fhit. The recently determined crystal structure of the NitFhit fusion protein has not made the situation much clearer (Pace *et al.*, 2000).

It remains unclear whether such approaches are truly general, in particular for proteins such as receptors that span different cellular compartments. For example, some receptor tyrosine kinases contain a kringle domain in their extracellular regions. Would such protocols predict common functions for intracellular tyrosine kinases and extracellular kringle-containing proteins, such as those of the blood coagulation pathway? Nevertheless, it is apparent that considerable functional constraints exist for domains to co-occur and that domain combinations are often very limited.

IV. Genome-Wide Analysis: New Quality in Domain Research

The availability of completely sequenced genomes offers the possibility of studying gene, protein, and domain evolution at the organismal level. Domain-based analyses confer several benefits to these studies. Not only can annotations be improved in various ways, but also the evolution of the many multidomain protein families can be traced. The latter requires a careful distinction between homology detection and orthology identification. Orthology identification is crucial for many approaches in comparative genome analysis and should be carried out at both the protein level and the domain level. As the postgenome era of cellular organisms is only a few years old, comparative analysis of entire genomes is in its early stages: domain analysis in this context has only recently been applied in a very rudimentary form. We expect this to change in the near future; the topics below represent only a few examples to indicate the different levels that can be exploited.

A. Domains as a Tool to Aid Gene Prediction

Domain analysis is particularly important for gene identification and annotation, as well as for the detection of misassembled genes. For example, a *C. elegans* gene (C35B8.2), predicted by the Genome Sequencing Project to encode the orthologue of human Vav, contains a different domain architecture from Vav (Fig. 5a). Further analysis of the genome sequence reveals that this gene has been misassembled and that a corrected sequence does indeed contain all the domains seen for its human orthologue. Intriguingly, the *Drosophila* orthologue (CG7893) of this gene appears to be missing the more N-terminal SH3 domain with apparently no room in the genomic sequence to accommodate it (Dekel *et al.*, 2000). Similar phenomena may not be associated with errors,

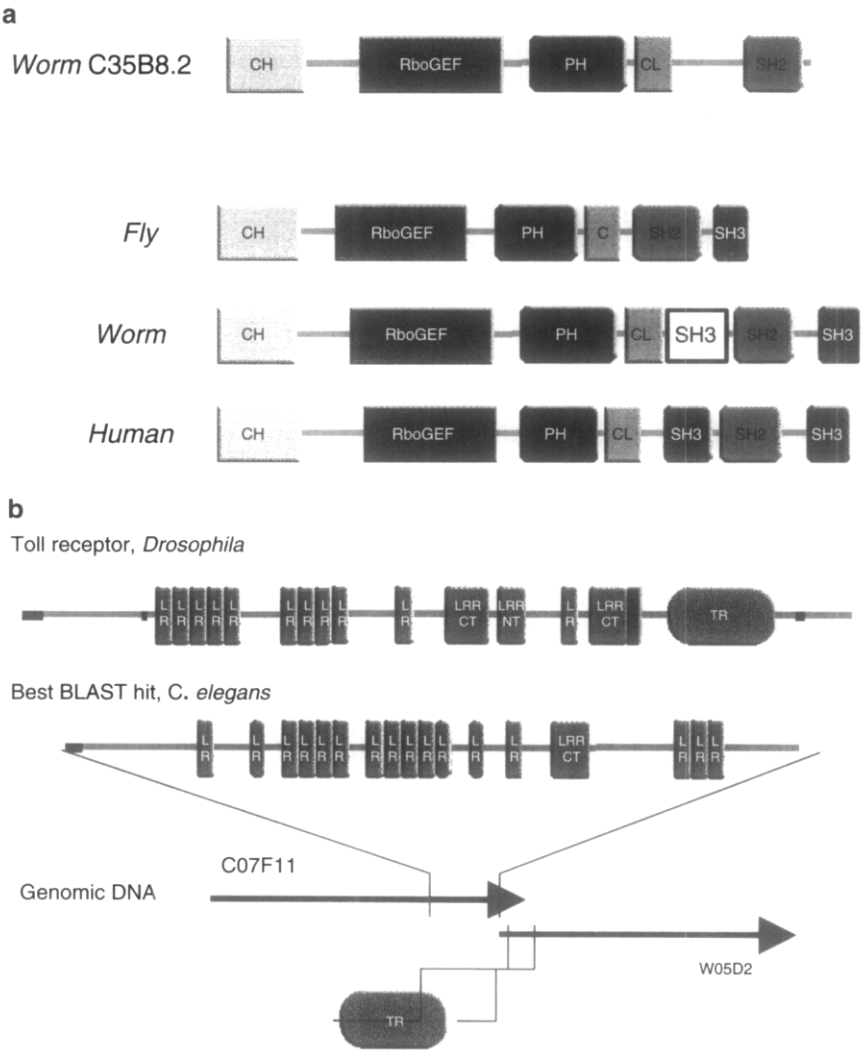

FIG. 5. (a) Vav like proteins in *C. elegans*, *D. melanogaster* (SP-TREMBL Q9NHV9), and humans (P15498). The predicted *C. elegans* protein C35B8.2 lacks a C-terminal **SH3** domain that is apparent in the genomic sequence. The *Drosophila* protein lacks the first **SH3** domain, and the absence of genomic DNA of sufficient length suggests that this is not a failure to predict the domain. The first **SH3** domain in *C. elegans* is very divergent and not predicted by SMART. (b) The *Drosophila* Toll receptor (P08953) was used to identify the best match in the predicted proteins of *C. elegans* (wormpep id C07F11.1). This predicted protein shows the extracellular leucine-rich repeats (**LRR**) and associated domains, but does not contain a **TIR** domain. The protein is found at the end of a genomic DNA clone (C07F11). Analysis of the following clone (W05D2) using genewise revealed the presence of a previously unpredicted TIR domain.

but with the complexities of alternatively spliced gene products (Black, 2000).

Understanding the domain structure of proteins and an accurate gene prediction process go hand in hand. Gene prediction is greatly assisted by using homology information (i.e., sequence similarity), but in the case of multidomain proteins, similarity may be difficult to detect or may not extend over the full length of the protein. The **TIR** domain is found in a large family of Toll-like receptors in *Drosophila* and in humans (Fig. 5b). These proteins are believed to be involved in the innate immune response (Aderem and Ulevitch, 2000). The predicted *C. elegans* protein set (wormpep, http://www.sanger.ac.uk/Projects/C_elegans/wormpep/) at the date of this writing contained only a single example of a **TIR** domain in a molecular context different from that of the Toll-like receptors (Aravind *et al.*, 1999). However, by identifying the *C. elegans* peptide with the best match to the extracellular portion of a *Drosophila* toll receptor and using sensitive software to identify matches between genomic DNA and protein HMMs (genewise, http://www.sanger.ac.uk/Software/Wise2), we were able to confirm the presence of a previously unpredicted **TIR** domain (Fig. 5b). The details are important. The value of model organisms lies in their increased simplicity: in this case, only a single Toll-like receptor is present in *C. elegans* compared to multiple copies in *Drosophila* and humans. Thus, the phenotype (if any) of knocking out the *C. elegans* gene is unlikely to be obscured by functional redundancy of close copies.

To understand the evolution of function and of multidomain proteins, artifacts arising due to sequence or annotation errors must be distinguished from the genuine products of domain shuffling or acquisition. For example, the *C. elegans* sequence most similar to that of human Tollip is F25H2.1 (Fig. 4), yet the worm sequence contains an extra domain at its C-terminal end. It remains a matter of conjecture whether it represents an aberrant gene fusion of two distinct genes or whether the worm sequence is genuine. If genuine, then several questions arise. Are these human and *C. elegans* genes functionally equivalent? Can the function of one be inferred from the other? And, can one decide on a definition of whether such genes are "equivalent" or not?

B. Orthology and Paralogy

Equivalent genes in different species have been named orthologues (Fitch, 1970). The value of discriminating between genes that arose from speciation events (orthologues) and genes that arose from intragenomic

duplication events (paralogues) was not fully recognized until the mid-1990s when the first microbial genomes were sequenced and the extent of intragenomic duplication became apparent. For function prediction, the distinction is crucial: different members of multigene families may possess distinct functions. Indeed, for paralogues to persist in a genome, it is likely that there must be some distinction in function, with subtle changes in either expression or specificity (Lynch and Conery, 2000, and references therein). Consequently, only orthologues should be used as excellent predictors for the transfer of functional information. In practice, eukaryotic paralogues are often useful in accurately predicting function. This is supported by many experimental observations of eukaryotic paralogues possessing overlapping functions (for example, Manley and Capecchi, 1997; Teglund *et al.*, 1998; Tinsley *et al.*, 1998).

Although the concept of orthologues is clear, there are no foolproof methods for their identification. One working definition of the orthologues, O_A and O_B, of two genomes A and B, are the genes for which O_A is the most similar to O_B in a search of B, and O_B is the most similar to O_A in a search of A. However, in many instances this definition is too simplistic. This is because there are often not one-to-one, but many-to-many relationships between related genes in different organisms arising from distinct gene duplication events in separate evolutionary lineages. Another complication arises from the fusion, fission, deletion, or shuffling of domains. We suggest that the concept of orthology be applicable to *both* domains *and* proteins in a similar manner to the current application of homologous domains and proteins. Thus the case might arise whereby orthologous domains in different species co-occur in different nonorthologous multidomain contexts. In this scenario, it is expected that the proteins differ in function despite their constituent orthologous domains possessing equivalent functions.

Issues relating to orthology and paralogy are crucial for genome comparison and annotation. This is particularly true for the genomes of multicellular organisms since, with their larger genomes, they encode many duplicated genes, a high proportion of which encode multidomain proteins. However, due to the methodological problems described above, domain analysis in such situations is currently restricted to homology identification and domain counting approaches (The *C. elegans* Sequencing Consortium, 1998; Rubin *et al.*, 2000; Lander *et al.*, 2001; Ponting *et al.*, 1999). These have proved to be important in highlighting the frequency of lineage-specific domain expansions (Fig. 4, and below). However, it is clear that advances in orthology/paralogy identification are required if more complex orthologous relationships are to be recognized. The use of domain analysis can help resolve more complex scenarios as outlined below.

C. Comparative Analysis and Evolution of Function

As complete genomes become available and orthologues are assigned, it becomes possible to trace the evolution of the domain structure of proteins. Processes of domain loss or domain gain must be reconciled with what is known of species phylogeny. Figure 6a illustrates a potential case of domain gain. The UPF0034 domain from Pfam (Bateman *et al.*, 2000) appears fused with a double stranded RNA-binding motif (**DSRM**) in certain metazoan (human, *Drosophila*, and *C. elegans*) proteins. However, the apparent yeast orthologue of these proteins contains no **DSRM** domain. This case again illustrates the potential values and difficulties of using domain fusion to predict function. The UPF0034 domain is poorly characterized, but is probably a phosphate-binding $(\beta/\alpha)_8$ barrel (Copley and Bork, 2000). The fusion with a **DSRM** domain suggests that its substrate may be RNA, providing a testable hypothesis where none was present before. Intriguingly, the yeast protein possesses a weak ability to suppress a defect in faulty mitochondrial $tRNA^{Asp}$ processing (Rinaldi *et al.*, 1997).

Within the eukaryotic crown group, Viridiplantae (i.e., plants) are believed to be less closely related to the metazoans than fungi (e.g., Baldauf *et al.*, 2000). Thus if a domain combination is shared between plants and metazoa, but not with fungi, the most parsimonious explanation is that it has been altered in the lineage leading to fungi. Figure 6b shows such a situation, with the potential loss of **SANT** domains from an ancestral gene encoding both **DnaJ** and **SANT** domains.

In cases where the organismal phylogeny is unknown, it is possible that shared derived domain combinations can be used as synapomorphies (i.e., if it is assumed that the evolution of that particular domain structure is monophyletic) to resolve the branching order of the species. It has been argued that *Drosophila* and *C. elegans* can be united in a clade of molting animals, known as *ecdysozoa* (Aguinaldo *et al.*, 1997). However, some evidence from the domain structure of equivalent proteins found in humans *Drosophila*, and *C. elegans* is in conflict with this hypothesis (Lander *et al.*, 2001). In the coming years, with the arrival of more complete genome sequences from crown-group eukaryotes, such questions will receive close attention.

Evolution reworks the domain structure of proteins, modifying them by the addition and deletion of domains. A parallel evolutionary theme highlighted by comparative analysis is the expansion or contraction of particular domain families within phylogenetic lineages. Analysis of the complete genomes of *C. elegans* and *D. melanogaster* reveals striking examples of expansion of different domain families, particularly in the complement of extracellular proteins. As has previously been noted

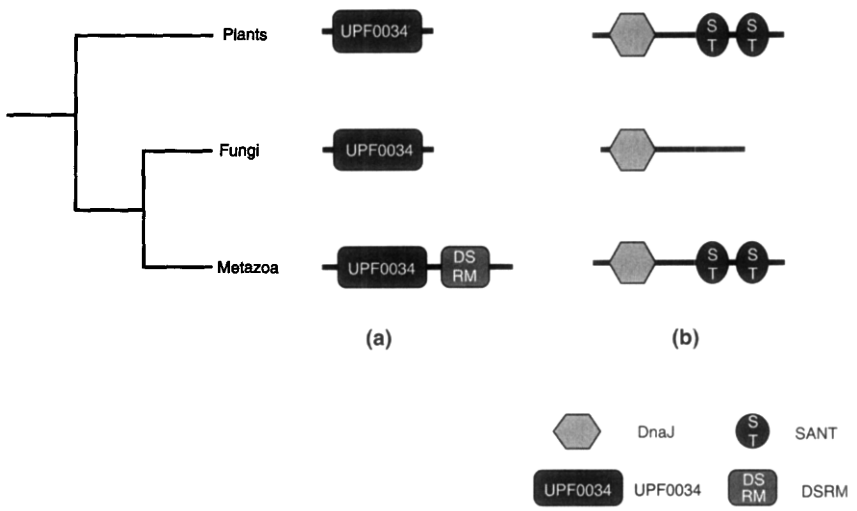

FIG. 6. (a) Fusion of double-stranded RNA-binding Motif (**DSRM**) with UPF0034, an uncharacterized domain from PFAM (Accession No. PF01207). The sequences shown are predicted to be orthologous. Sequence accession numbers used: SP–TREMBL: *C. elegans*, Q9XWJ9; *D. melanogaster*, Q9VY45; *S. cerevisiae*, P53720; *H. sapiens*, Q9NX74. (b) Potential loss of **SANT** domains. Sequences: *S. cerevisiae*, P32527; *H. sapiens*, 060415; *D. melanogaster*, Q9VP77; *C. elegans*, Q94216; *Arabidopsis thaliana*, Q9SS16, Q9LHS5. The orthologous grouping of the common domains (i.e., UPF0034 or **DnaJ**) was, in both cases, supported by bootstrapping. However, the branching within the groups is less well supported. Accordingly, the tree represents the standard species phylogeny. Two additional explanations could also account for such apparent rearrangements of domain architecture: (1) invoking an ancestor with both domain structures, followed by multiple gene losses, and (2) aberrant protein prediction from DNA sequences.

(Rubin *et al.*, 2000), the classical trypsin-like serine protease family (**Tryp_SPc**) is greatly expanded in *Drosophila* compared to *C. elegans*. This expanded family includes *easter*, *snake*, and *gastrulation defective* from the Toll pathway, crucial to dorsoventral patterning in *Drosophila*. Despite the importance of the pathway in *Drosophila* development, neither these serine proteases nor other components of the pathway have counterparts in *C. elegans*. Rapid expansion of catalytic sequence families such as this may be partially explained by the evolution and duplication of protein cascades, where similar proteins act on one another to amplify an initial signal (Caffrey *et al.*, 1999). In addition to expansion, these domains provide an example of another general theme in domain evolution: an arrangement of domains unique to a particular phylogenetic lineage. Although they were initially reported to be single-domain proteins (Rubin *et al.*, 2000), close inspection reveals that many **Tryp_SPc**

domain-containing proteins (including *snake* and *easter*) have a conserved motif at the N-terminus (**CLIP**) that appears to be unique to the arthropods (Smith and De Lotto, 1992). A similar motif is also found in four copies in the *Drosophila* masquerade protein. Interestingly, here the serine protease domain is lacking essential catalytic residues and so is unlikely to function as an enzyme (Murugasu-Oei *et al.*, 1995).

In contrast to this relatively simple domain structure, mammalian serine proteases exhibit a greater architectural complexity (despite a lower number of proteases being encoded in the genome). This may be accounted for by far greater demands for regulational complexity related to their roles in, for instance, the regulation of digestion and blood clotting cascades (Fig. 7).

Interesting as this may be, and shedding light on some of the most fundamental aspects of molecular evolution, these analyses are, hopefully, only a beginning. The ultimate goal must be to integrate these studies with function, phenotypes, and population dynamics if we are to truly succeed in understanding the molecular nature of biological change.

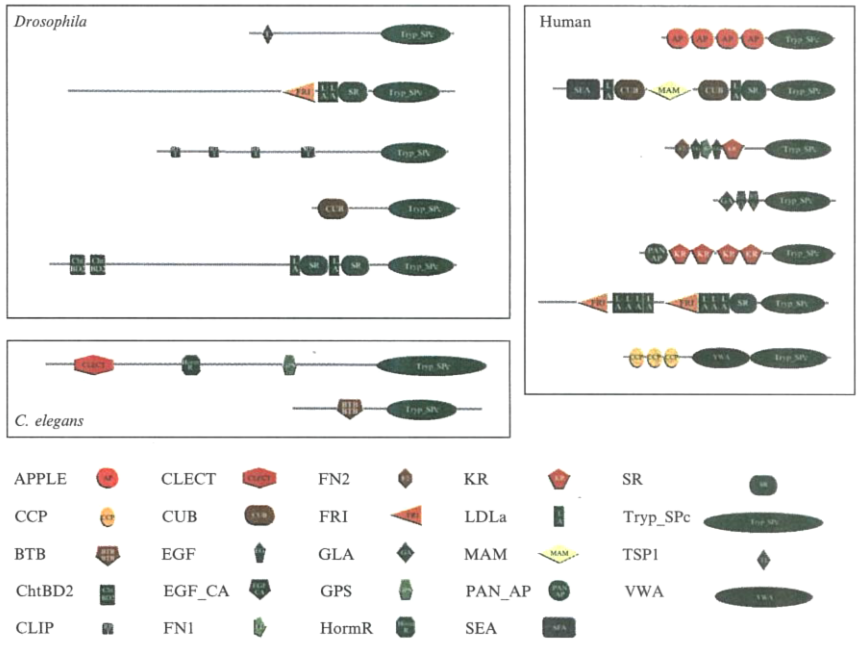

FIG. 7. Trypsin-like serine proteases in *C. elegans*, *D. melanogaster*, and humans. The figure shows illustrative examples of the distinct domain co-occurrences with the serine protease-like domains in each species. The key shows the SMART (http://smart.embl-heidelberg.de/) names of the co-occurring domains.

V. Conclusion

Domain identification and analysis are essential for understanding protein structure and evolution and provide the starting point from which comparative analysis of genomes can proceed. Although most of the widespread domain families have likely already been identified, we are far from having a full catalogue of all domains. Such a complete catalogue would include representatives of all sequence families, within the unifying conceptual framework of a structural (and evolutionary) hierarchy, through which the emergence of function could be traced with increasing levels of precision. By applying what is already known, we can begin a systematic classification and quantification of evolutionary events. Linking such computational analyses to the coming waves of data from functional genomics projects, where information is generated for thousands of proteins at a time (e.g., Gönzcy *et al.*, 2000; Fraser *et al.*, 2000), will stand us in good stead as we attempt to move ever closer to an understanding of our nature.

References

Adams, M. J., *et al.*, (2000). *Science* **287**, 2185–2195.
Aderem, A., and Ulevitch, R. J. (2000). *Nature* **406**, 782–787.
Aguinaldo, A. M., Turbeville, J. M., Linford, L. S., Rivera, M. C., Garey, J. R., Raff, R. A., and Lake, J. A. (1997). *Nature* **387**, 489–493.
Altschul, S. F., and Gish, W. (1996). *Methods Enzymol.* **266**, 460–480.
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.
Andrade, M. A., Ponting, C. P., Gibson, T. J., and Bork, P. (2000). *J. Mol. Biol.* **298**, 521–537.
Aravind, L., Dixit, V. M., and Koonin, E. V. (1999). *Trends Biochem. Sci.* **24**, 47–53.
Baldauf, S. L., Roger, A. J., Wenk-Siefert, I., and Doolittle, W. F. (2000). *Science* **290**, 972–977.
Baron, M., Norman, D. G., and Campbell, I. D. (1991). *Trends Biochem. Sci.* **16**, 13–17.
Bateman, A., Birney, E., Durbin, R., Eddy, S. R., Howe, K. L., and Sonnhammer, E. L. (2000). *Nucleic Acids Res.* **28**, 263–266.
Black, D. L. (2000). *Cell* **103**, 367–370.
Blomberg, N., and Nilges, M. (1997). *Fold. Des.* **2**, 343–355.
Bork, P. (1992). *Curr. Opin. Struct. Biol.* **2**, 413–421.
Bork, P. (1996). *Matrix Biol.* **15**, 311–312.
Bork, P., and Bairoch, A. (1995). *Trends Biochem. Sci.* **20** (Poster Supplement C02).
Bork, P., and Koonin, E. V. (1996). *Curr. Opin. Struct. Biol.* **6**, 366–376.
Bork, P., Schultz, J., and Ponting, C. P. (1997). *Trends Biochem. Sci.* **22**, 296–298.
Bucher, P., Karplus, K., Moeri, N., and Hofmann, K. (1996). *Comput. Chem.* **20**, 3–23.
Caffrey, D. R., O'Neill, L. A., and Shields, D. C. (1999). *J. Mol. Evol.* **49**, 567–582.
Callebaut, I., Cossart, P., and Dehoux, P. (1998). *FEBS Lett.* **441**, 181–185.
Casari, G., Sander, C., and Valencia, A. (1995). *Nat. Struct. Biol.* **2**, 171–178.
C. elegans Sequencing Consortium (1998). *Science* **282**, 2012–2018.

Copley, R. R., and Bork, P. (2000). *J. Mol. Biol.* **303**, 627–641.
Corpet, F., Servant, F., Gouzy, J., and Kahn, D. (2000). *Nucleic Acids Res.* **28**, 267–269.
Dekel, I., Russek, N., Jones, T., Mortin, M. A., and Katzav, S. (2000). *FEBS Lett.* **472**, 99–104.
Deshimaru, M., Ogawa, T., Nakashima, K., Nobuhisa, I., Chijiwa, T., Shimohigashi, Y., Fukumaki, Y., Niwa, M., Yamashina, I., Hattori, S., and Ohno, M. (1996). *FEBS Lett.* **397**, 83–88.
Doerks, T., Copley, R., Schultz, J., Ponting, C. P., and Bork, P. (2002). *Genome Res.* **12**, 47–56.
Doolittle, R. F. (1985). *Trends Biochem. Sci.* **10**, 233–237.
Enright, A. J., Iliopoulous, I., Kyripides, N. C., and Ouzounis, C. A. (1999). *Nature* **402**, 86–90.
Espanel, X., and Sudol, M. (1999). *J. Biol. Chem.* **274**, 17284–17289.
Fitch, W. M. (1970). *Syst. Zool.* **19**, 99–113.
Fraser, A. G., Kamath, R. S., Zipperlen, P., Martinez-Campos, M., Sohrmann, M., and Ahringer, J. (2000). *Nature* **408**, 325–330.
Gilbert, W. (1978). *Nature* **271**, 501.
Gilbert, W. (1985). *Science* **228**, 823–824.
Go, M. (1983). *Proc. Natl. Acad. Sci. USA* **80**, 1964–1968.
Gönczy, P., Echeverri, C., Oegema, K., Coulson, A., Jones, S. J. M., Copley, R. R., Duperon, J., Oegema, J., Brehm, M., Cassin, E., Hannak, E., Kirkham, M., Pichler, S., Flohrs, K., Goesson, A., Leidel, S., Alleaume, A-M., Martin, C., Ozlu, N., Bork, P., and Hyman, A. A. (2000). *Nature* **408**, 331–336.
Gouzy, J., Corpet, F., and Kahn, D. (1999). *Comput. Chem.* **23**, 333–340.
Gribskov, M., McLachlan, A. D., and Eisenberg, D. (1987). *Proc. Natl. Acad. Sci. USA* **84**, 4355–4358.
Hannenhalli, S. S., and Russell, R. B. (2000). *J. Mol. Biol.* **303**, 61–76.
Heger, A., and Holm, L. (2000). *Prog. Biophys. Mol. Biol.* **73**, 321–337.
Hofmann, K., Bucher, P., Falquet, L., and Bairoch, A. (1999). *Nucleic Acids Res.* **27**, 215–219.
Isakoff, S. J., Cardozo, T., Andreev, J., Li, Z., Ferguson, K. M., Abagyan, R., Lemmon, M. A., Aronheim, A., and Skolnik, E. Y. (1998). *EMBO J.* **17**, 5374–5387.
Janin, J., and Chothia, C. (1985). *Methods Enzymol.* **115**, 420–430.
Kalhammer, G., Bahler, M., Schmitz, F., Jockel, J., and Block, C. (1997). *FEBS Lett.* **414**, 599–602.
Karlin, S., and Altschul, S. F. (1990). *Proc. Natl. Acad. Sci. USA* **87**, 2264–2268.
Kimber, M. S., Nachman, J., Cunningham, A. M., Gish, G. D., Pawson, T., and Pai, E. F. (2000). *Mol. Cell* **5**, 1043–1049.
Lander, E. S., *et al.* (2001). *Nature* **409**, 860–921.
Lichtarge, O., Bourne, H. R., and Cohen, F. E. (1996). *J. Mol. Biol.* **257**, 342–358.
Lo Conte, L., Ailey, B., Hubbard, T. J., Brenner, S. E., Murzin, A. G., and Chothia, C. (2000). *Nucleic Acids Res.* **28**, 257–259.
Lupas, A. N., Ponting, C. P., and Russell, R. B. (2001), *J. Struct. Biol.* **134**, 191–203.
Lynch, M., and Conery, J. S. (2000). *Science* **290**, 1151–1155.
Makarova, K. S., and Grishin, N. V. (1999). *J. Mol. Biol.* **292**, 11–17.
Manley, N. R., and Capecchi, M. R. (1997). *Dev. Biol.* **192**, 274–288.
Marcotte, E. M., Pellegrini, M., Ng, H. L., Rice, D. W., Yeates, T. O., and Eisenberg, D. (1999a). *Science* **285**, 751–753.
Marcotte, E. M., Pellegrini, M., Thompson, M. J., Yeates, T. O., and Eisenberg, D. (1999b). *Nature* **402**, 83–86.
Mott, R. (2000). *J. Mol. Biol.* **300**, 649–659.
Murugasu-Oei, B., Rodrigues, V., Yang, X., and Chia, W. (1995). *Genes Dev.* **9**, 139–154.
Needleman, S. B., and Wunsch, C. D. (1970). *J. Mol. Biol.* **48**, 443–453.
Pace, H. C., Hodawadekar, S. C., Dragenscu, A., Huang, J., Bieganowski, P., Pekarsky, Y., Croce, C. M., and Brenner, C. (2000). *Curr. Biol.* **10**, 907–917.

Patthy, L. (1985). *Cell* **41**, 657–663.
Patthy, L. (1987). *FEBS Lett.* **214**, 1–7.
Patthy, L., Trexler, M., Vali, Z., Banyai, L., and Varadi, A. (1984). *FEBS Lett.* **171**, 131–136.
Pearl, F. M., Lee, D., Bray, J. E., Silltoe, I., Todd, A. E., Harrison, A. P., Thornton, J. M., and Orengo, C. A. (2000). *Nucleic Acids Res.* **28**, 277–282.
Pearson, W. R. (1998). *J. Mol. Biol.* **276**, 71–84.
Pekarsky, Y., Campiglio, M., Siparashvili, Z., Druck, T., Sedkov, Y., Tillib, S., Draganescu, A., Wermuth, P., Rothman, J. H., Huebner, K., Buchberg, A. M., Mazo, A., Brenner, C., and Croce, C. M. (1998). *Proc. Natl. Acad. Sci. USA* **95**, 8744–8749.
Ponting, C. P. (2000). *Biochem. J.* **351**, 527–535.
Ponting, C. P., Aravind, L., Schultz, J., Bork, P., and Koonin, E. V. (1999). *J. Mol. Biol.* **289**, 729–745.
Ponting, C. P., and Russell, R. B. (2000). *J. Mol. Biol.* **302**, 1041–1047.
Rinaldi, T., Lande, R., Bolotin-Fukuhara, M., and Frontali, L. (1997). *Curr. Genet.* **31**, 494–496.
Rossmann, M. G., Moras, D., and Olsen, K. W. (1974). *Nature* **250**, 194–199.
Rubin, G. M., *et al.* (2000). *Science* **287**, 2204–2215.
Russell, R. B., and Ponting, C. P. (1998). *Curr. Opin. Struct. Biol.* **8**, 364–371.
Schultz, J., Milpetz, F., Bork, P., and Ponting, C. P. (1998). *Proc. Natl. Acad. Sci. USA* **95**, 5857–5864.
Schultz, J., Copley, R. R., Doerks, T., Ponting, C. P., and Bork, P. (2000). *Nucleic Acids Res.* **28**, 231–234.
Shapiro, L., and Scherer, P. E. (1998). *Curr. Biol.* **8**, 335–338.
Sjolander, K. (1998). *I.S.M.B.* **6**, 165–174.
Smith, C. L., and De Lotto, R. (1992). *Protein Sci.* **1**, 1225–1226.
Smith, T. F., and Waterman, M. S. (1981). *J. Mol. Biol.* **147**, 195–197.
Soisson, S. M., Nimnual, A. S., Uy, M., Bar-Sagi, D., and Kuriyan, J. (1998). *Cell* **95**, 259–268.
Teglund, S., McKay, C., Schuetz, E., van Deursen, J. M., Stravopodis, D., Wang, D., Brown, M., Bodner, S., Grosveld, G., and Ihle, J. N. (1998). *Cell* **93**, 841–850.
Tinsley, J., Deconinck, N., Fisher, R., Khan, D., Phelps, S., Gillis, J. M., and Davies, K. (1998). *Nat. Med.* **4**, 1441–1444.
Tordai, H., Banyai, L., and Patthy, L. (1999). *FEBS Lett.* **461**, 63–67.
Tsujishita, Y., and Hurley, J. H. (2000). *Nat. Struct. Biol.* **7**, 408–414.
Wilson, C. A., Kreychman, J., and Gerstein, M. (2000). *J. Mol. Biol.* **297**, 233–249.