Protein families in multicellular organisms

Richard R Copley*, Jörg Schultz*, Chris P Ponting† and Peer Bork*‡

The complete sequence of the nematode worm *Caenorhabditis elegans* contains the genetic machinery that is required to undertake the core biological processes of single cells. However, the genome also encodes proteins that are associated with multicellularity, as well as others that are lineage-specific expansions of phylogenetically widespread families and yet more that are absent in non-nematodes. Ongoing analysis is beginning to illuminate the similarities and differences among human proteins and proteins that are encoded by the genomes of the multicellular worm and the unicellular yeast, and will be essential in determining the reliability of transferring experimental data among phylogenetically distant species.

Addresses
*Biocomputing, European Molecular Biology Laboratory, Meyerhofstrasse 1, 69012 Heidelberg, Germany
†National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Building 38A, Bethesda, MD 20894, USA
‡e-mail: bork@embl-heidelberg.de

Current Opinion in Structural Biology 1999, **9**:408–415

http://biomednet.com/elecref/0959440X00900408

Abbreviations

EGF	epidermal growth factor
FN3	fibronectin type 3
PH	pleckstrin homology
PX	phox homologous
SH	Src homology

Introduction

Over the past five years, complete genome sequences have allowed the analysis of the basic metabolic and replication machinery of archaeal, eukaryotic and bacterial single-cell life forms [1,2]. With the essentially complete genome of the eukaryotic multicellular organism *C. elegans* now available [3••], we can begin to address the questions that are central to our understanding of the evolution of multicellular life. To what extent are the obvious differences between multicelled and single-celled life reflected in differences among the protein complements encoded by their genomes? Can proteins that are specific to multicellular organisms be identified? How are human proteins related to those of *C. elegans*? Is there a molecular profile of multicellularity that is conserved between animals and plants? Here, we confine our discussion of multicellularity to the plant and animal kingdoms and, as far greater data are available for metazoa than for plants, much of the following will focus on metazoa.

The genomes of multicellular organisms

The 19,099 proteins encoded by the complete genome of *C. elegans* go some way to filling the vacuum in our knowledge of multicellular organisms [3••]. Even so, their true value will become all the more apparent as other genome projects draw to their conclusions. The complete genome sequence of the fruit fly *Drosophila melanogaster* (with an estimated 8000 to 15,000 genes) is likely to be available by the end of 2001 [4], but probably earlier [5], and 90% of the human genome is likely to be available by Spring 2000. Coupling the wealth of developmental data available for *C. elegans* and *D. melanogaster* to their complete genome sequences will provide invaluable insights into the evolution of development and, hopefully, guide the interpretation of developmental processes in *Homo sapiens* and other metazoa.

In order to determine which proteins are specific to multicellular life, a comparison with single-celled organisms must be made. At present, the only available representative of a completely sequenced unicellular eukaryotic genome is that of the fungus *Saccharomyces cerevisiae*. This should soon be joined by the complete genome of a second fungus, *Schizosaccharomyces pombe*, but the complete genomes of nonfungal, early branching, single-celled eukaryotes would be of immense value. Equally, as multicellularity is believed to have evolved separately in animals and plants, it will be instructive to contrast metazoan genomes with those of plants. A project is underway to sequence the complete genome of *Arabidopsis thaliana*, with an estimated 20,000 genes [6].

A powerful method for comparing the genomes of single-celled organisms is the identification of orthologous pairs of proteins. These are homologues from different organisms that are related by a past speciation event, rather than being products of an intragenome duplication event. The value of orthologous relationships is that they enable functional data to be transferred between species. Orthologue identification and, hence, function prediction are, however, complicated by several factors, including sequencing error, incorrect exon prediction, non-equivalent multidomain architectures and the consequences of past genome duplication events [7]. Two independent genome duplication events are believed to have taken place in the chordate lineage [8], resulting in up to four orthologues in mammals for each *C. elegans* protein. For some species, the situation is yet more complex; the zebrafish *Danio rerio* and the teleost lineage are believed to have undergone an additional round of genome duplication [9,10]. Thus, each *C. elegans* protein might have a maximum of eight orthologues in *D. rerio*. Subsequent local duplication of genes may increase numbers further. An intriguing consequence of genome duplication is that the functions of paralogous genes are wholly redundant at the time of duplication and are partially redundant thenceforth. Much discussion has centred on the apparent

Table 1

Pairwise percentage identities of the 20 most conserved proteins present in *C. elegans*, *H. sapiens* and *S. cerevisiae*, as determined by reciprocal BLAST searches*.

	Description	Pairwise percentage identity		
		Worm/Human	Worm/Yeast	Yeast/Human
1	H4 histone	99	91	92
2	H3.3 histone	99	89	90
3	Actin B	98	88	89
4	Ubiquitin	98	95	96
5	Calmodulin	96	59	58
6	Tubulin 2	94	75	76
7	Ubiquitin-conjugating enzyme UBC4	93	80	80
8	Clathrin coat associate protein	93	48	48
9	Tubulin	93	73	74
10	ADP ribosylation factor 1	93	77	77
11	Dynein light-chain 1	92	51	50
12	GTP-binding nuclear protein RAN	89	82	81
13	Ser/Thr protein phosphatase PP1γ	89	84	85
14	Ser/Thr protein phosphatase PP2β	89	74	76
15	Ubiquitin-conjugating enzyme UBE2N	88	67	70
16	Histone H2A.Z	88	69	69
17	Histone H2A.2	87	79	76
18	DIM1P homologue	86	61	65
19	G25K GTP-binding protein	86	76	80
20	40S ribosomal protein S15A	86	76	77

*The yeast protein set was obtained from ftp://genome-ftp.stanford.edu/pub/yeast/yeast_protein/. The 19,099 *C. elegans* sequences were taken from the October 1998 set at http://www.sanger.ac.uk/Projects/C_elegans/Science98/. The human proteins were obtained by filtering SPTREMBL at the 97% identity level, leaving 11,709 sequences.

paradox of paralogue persistence, as functionally redundant genes might be expected to be deleted relatively rapidly from the genome [11,12•].

Conserved proteins

There are many biological processes that are common to the majority of cellular life and these are reflected in the sets of proteins that are conserved across very large phylogenetic distances [13,14•] (see Table 1 for examples of well-conserved proteins). A comparative analysis of all of the predicted proteins of *C. elegans* and *S. cerevisiae* revealed a set of biological functions that is shared between the two organisms [14•]. Many one-to-one orthologous relationships are detectable and the majority of these can be assigned a function corresponding to core biological processes (e.g. intermediary metabolism and DNA/RNA metabolism) [14•]. Mushegian *et al.* [13] searched for orthologous proteins encoded by the available data from the *H. sapiens*, *D. melanogaster*, *C. elegans* and *S. cerevisiae* genomes. The incomplete datasets for all but the *S. cerevisiae* genome made the unequivocal inference of orthology difficult. A total of 42 quartets of candidate orthologues were found, the majority of which were enzymes or were involved in DNA replication. This number will increase, as more data become available from *D. melanogaster* and *H. sapiens*. The value of such studies lies in the demonstration that, even over large phylogenetic distances, useful functional information can be inferred and the proteins responsible for basic biological processes can be identified from sequence data alone. Others have suggested that more caution is advisable [15].

Signalling

In order to coordinate their functions, the cells of a multicellular organism must maintain a constant flux of intercellular communication. The proteins involved in these signalling processes are often multidomain in character and contain components drawn from a large repertoire of eukaryotic signalling domains. Consequently, it is thought that the acquisition of new functions during the evolution of multicellular eukaryotic signalling has been reliant more upon the expansion of existing domain families and their association in novel combinations, rather than on the genesis of novel domain types.

The majority of signalling regulatory domains are found only in eukaryotes and appear in considerable numbers of eukaryotic lineage-specific multidomain contexts. This and the considerable divergence in their sequences ensure that such domains are not straightforward to detect using simple database searching methods. Fortunately, automated detection methods, such as the SMART, PROSITE and PFAM [16•–18•,19] systems, now offer ease of domain identification via the World Wide Web. The SMART system, with which we are most familiar, provides a simple interface for sensitive searching of protein sequences for intracellular and extracellular signalling domains and contains hyperlinks both to structural and functional data and also to literature that is relevant to each domain.

Table 2

Analysis of signalling domain numbers in *H. sapiens*, *C. elegans* and *S. cerevisiae* datasets (see Table 1), performed using the SMART system.

Table 2a. Intracellular SMART domains present in *H. sapiens* and *C. elegans*, but not *S. cerevisiae*.

Domain	Description	Human	Worm
PTB	Phosphotyrosine binding	28	11
B41	Band 4.1 homologues	24	16
DEATH	Involved in apoptosis	22	9
CASc	Caspases	14	3
CARD	Caspase recruitment domain	12	2
ZU5	In ZO1 and netrin receptors	5	6
DAX	In Dishevelled and axin	5	3
MyTH4	In myosin and kinesin tails	3	4

Table 2b. Predominantly extracellular domains that have more than 50 copies in *C. elegans*.

Domain	Description	Human	Worm	Yeast
IG	Immunoglobulin	1078	410	0
CLECT	C-type lectin	73	313	0
FN3	Fibronectin type 3	587	213	2
ShKT	ShK toxin	1	259	0
LDLa	Low-density lipoprotein receptor class A	145	147	1
KU	BPTI/Kunitz family of protease inhibitors	20	139	0
TSP1	Thrombospondin type 1 repeats	104	119	0
CA	Cadherin repeats	284	118	0
CUB	In developmentally regulated proteins	65	94	0
CCP	Short complement-like repeat (SUSHI)	207	80	0

Intracellular signalling

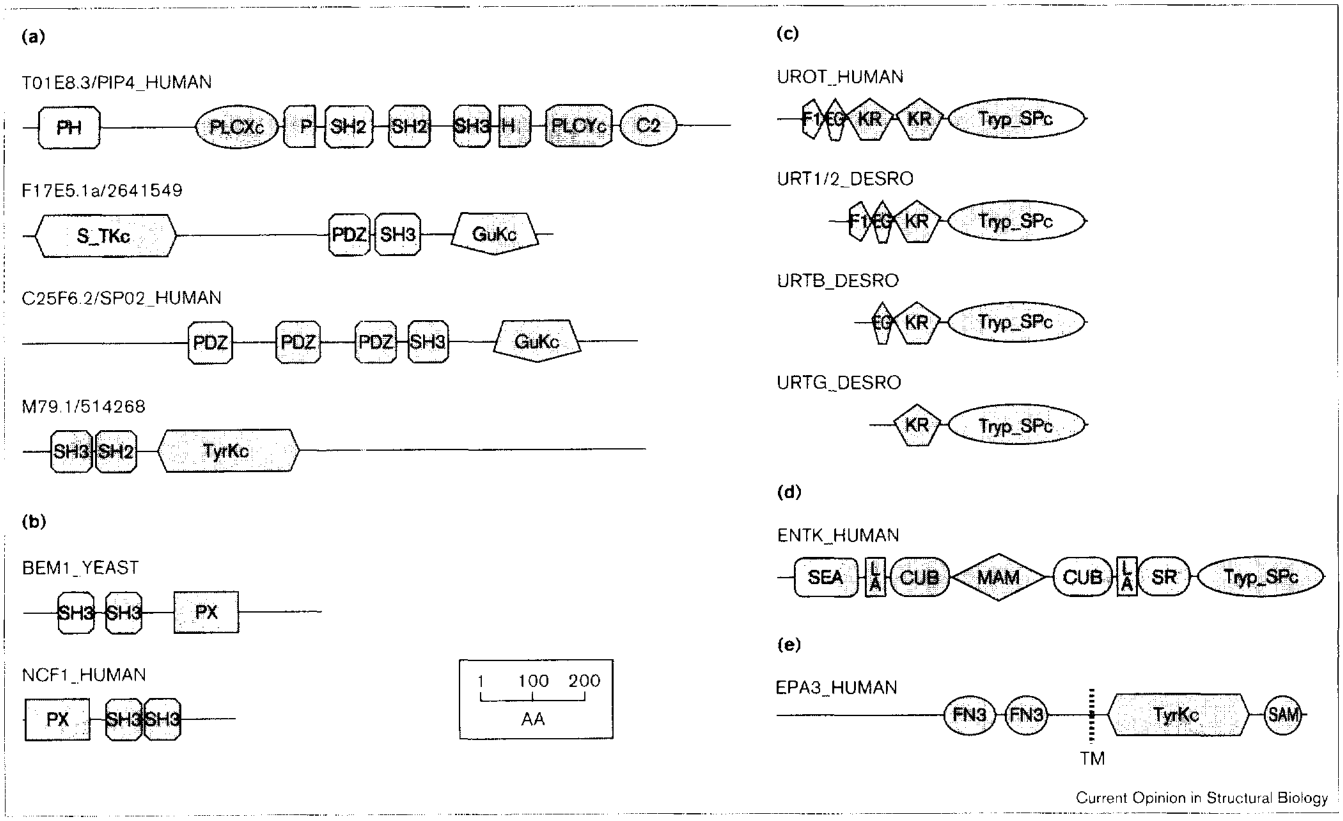
The majority of the intracellular signalling domains in the SMART database are found in both *S. cerevisiae* and *C. elegans* and, presumably, in most eukaryotes. Indeed, the majority of signalling enzymes, including the protein kinase catalytic domain [20], appear to have originated prior to the last common ancestor of eukaryotes, bacteria and archaea. By contrast, a minority of eukaryotic-type signalling regulatory domains are detectable in prokaryotes (CP Ponting, L Aravind, J Schultz, P Bork, EV Koonin, unpublished data).

Intracellular signalling domains that are absent in yeast, but present in worm are recorded in Table 2a. The presence or absence of some domains is related to multicellularity. The most obvious of these are families of domains that control apoptosis, namely the DEATH, CARD and caspase domains, which are all absent in yeast (another apoptotic signalling domain, DED, is found in mammals, but not in worm or yeast). Although apoptosis is a metazoan phenomenon, nonmetazoan cell death may sometimes display morphological characteristics that are reminiscent of apoptosis [21]. In fungi, this may be associated with the presence of a portion of the mammalian apoptotic machinery, as both *S. cerevisiae* and *S. pombe* contain inhibitors of apoptosis protein domains [22], meprin and TRAF-homology domains [23•] and BAG domains (SMART domains BIR, MATH and BAG). The latter family [24] includes the human 'silencer of death domains' protein [25] and hypothetical proteins in *C. elegans*, yeasts and *A. thaliana*.

Although caspases have not been detected in nonmetazoa, they have recently been identified as being members of wider superfamily of homologous cysteine endopeptidases, including legumain, clostripain and gingipains [23•,26]. As these proteins have a broad phylogenetic distribution, including plants, fungi and bacteria, a more ancient origin for these proteins is suggested. It would appear that an ancestral caspase-like molecule was recruited to a new function, with subsequent expansion of the family. No plant homologues of DEATH, CARD and DED have been reported. Plant resistance gene products have recently been shown to contain the NB-ARC domain, however, in common with the *C. elegans* CED-4 cell death adaptor protein [23•,27], and there is evidence of caspase-like activity in the hypersensitive response of plants to pathogens [28].

In addition to domains that are not found in yeast, many families are greatly expanded in *C. elegans*, implying that the domains have acquired new roles. For instance, there are five PDZ domains in yeast, 83 in worm and more than 150 in the available 10–15% of the total number of human protein sequences. This is consistent with a role for PDZ-domain-containing proteins in the organisation of higher order structures, for example, neurons and synapses [14•,29,30]. The protein kinase catalytic domain has acquired tyrosine kinase activity in metazoa and its presence in many receptors indicates that it is now crucial to multicellularity. There are no detectable tyrosine kinases in yeast, but there are at least 73 in worm and over 100 in the available human proteins. The reverse situation, that

Figure 1



Domain architectures of proteins described in the text. *C. elegans* accessions refer to the dataset described in the legend to Table 1. Other accessions are taken from SWISSPROT or the GenBank identifiers. The bar marked AA represents the length 200 amino acids. **(a)** Examples of intracellular domain architectures conserved between *C. elegans* and *H. sapiens* orthologues. PIP4_HUMAN, phospholipase Cγ; 2641549, calcium-dependent serine protein kinase; SPO2_HUMAN, synapse-associated protein 102; 514268, proto-oncogene tyrosine kinase. **(b)** An example of a multidomain architecture found in human, but not in *C. elegans*. A protein containing the same domains, although in a different arrangement, is found in *S. cerevisiae*. BEM-1_YEAST, BEM1 protein; NCF1_HUMAN, neutrophil cytosol factor 2. **(c)** Human plasminogen activator protein and vampire bat homologues. UROT_HUMAN, tissue plasminogen activator; URT1_DESRO, salivary plasminogen activator α; URTB_DESRO, salivary plasminogen activator β; URTG_DESRO, salivary plasminogen activator γ. **(d)** Human enterokinase domain structure. ENTK_HUMAN, enterokinase. **(e)** Domain architecture of an ephrin receptor (the extracellular domain [59] is not shown). EPA3_HUMAN, ephrin type-A receptor 3; TM, transmembrane domain. Abbreviations of SMART domains used in the figure. For further information, please see the web site http://coot.embl-heidelberg.de/SMART/. C2, protein kinase C conserved region 2 (CalB); CUB, domain first found in C1r, C1s, uEGF, and bone morphogenetic protein; F1, fibronectin type 1 domain; GuKc, guanylate kinase homologues; KR, kringle domain; LA, low-density lipoprotein receptor domain class A; MAM, domain in meprin, A5, receptor protein tyrosine phosphatase mu (and others); PDZ, domain present in PSD-95, Dlg, and ZO-1/2; PLCXc, phospholipase C, catalytic domain (part), domain X; PLCYc, phospholipase C, catalytic domain (part), domain Y; SAM, sterile alpha motif; SEA, domain found in sea urchin sperm protein, enterokinase, agrin; SR, scavenger receptor cysteine-rich; S_TKc, serine/threonine protein kinases, catalytic domain; Tryp_SPc, trypsin-like serine protease; TyrKc, tyrosine kinase, catalytic domain.

of proteins that are present in yeast being absent from metazoa, is observed with the histidine kinase catalytic domain. This domain is widespread in bacteria and archaea and is present in two copies in yeast; however, there are no detectable histidine kinases in human or *C. elegans*.

Given that the majority of intracellular signalling domains are present in yeast, what can be said about how these domains are assembled within multidomain proteins? Analysis of the context of intracellular signalling domains reveals profound differences between *S. cerevisiae* and *C. elegans*. The most commonly occurring intracellular signalling module is the protein kinase catalytic domain. In *S. cerevisiae*, 11 distinct SMART domains are found in kinase-containing proteins, whereas the equivalent figures for worm and human are 29 and 38, respectively. The domain structures of human kinases do not, however, represent the complete set from which all the others are drawn. Analysis of the incomplete *A. thaliana* genome reveals novel modular organisations of receptor kinases. For example, the extracellular portions of some plant receptor kinases contain EGF (epidermal growth factor) and bulb-type lectin domains [31], whereas others (namely F27L4.5 and F4P9.35) have been identified as containing pairs of lysis motif (SMART domain LysM) domains [32]. Given that searches for human receptor kinases have been extensive, it is unlikely that similar domain organisations will be found in metazoa.

Some intracellular signalling proteins with relatively complicated domain structures have been conserved in *C. elegans* and *H. sapiens*. For instance, *C. elegans* T01E8.3 contains all eight of the domains apparent in human phospholipase Cγ1, including a three-domain (SH2–SH2–SH3) insert into its C-terminal PH (pleckstrin homology) domain. More examples of such orthologues with multiple domains are illustrated in Figure 1a. Conversely, some intracellular signalling proteins are not conserved between these species. For example, there is no counterpart of human p47phox in worm or yeast, although the same domains (a PX domain [phox homologous domain, present in p47phox and p40phox] and two SH3 domains) are present in *S. cerevisiae* Bem1p, albeit in a different collinear arrangement (Figure 1b).

Extracellular signalling

The proportion of proteins predicted using the SignalP method [33] to be either partially extracellular or secreted is over twice as large in *C. elegans* (18%) as that in *S. cerevisiae* (8%) (J Schultz, unpublished data). This indicates a greatly expanded role for extracellular proteins in the nematode. In contrast to the situation with intracellular signalling domains, there appear to be many more distinct extracellular signalling domains in *C. elegans* than in *S. cerevisiae*. Even for those domains for which a copy is present, the family is often greatly expanded in *C. elegans* and human; for example, two copies of the FN3 (fibronectin type 3) domain are present in *S. cerevisiae*, compared with 213 in *C. elegans*. Table 2b lists SMART domains that have over 50 occurrences in worm, the majority of which do not appear to have yeast homologues.

Results from *C. elegans* indicate that it may contain many novel, extracellular, nematode-specific domains [34]. One domain family that is greatly expanded in worm, but not in human, is the set of homologues of the sea anemone ShK toxin. The structure of this domain is known [35] and an alignment of the homologous sequences is available from the SMART web site (identifier ShKT at http://coot.embl-heidelberg.de/SMART/). One hundred *C. elegans* proteins contain 259 copies of this domain, whereas only a single version is currently known in human sequences, namely in the metalloprotease MMP21/22.

Some multidomain extracellular proteins are conserved over large phylogenetic distances. For instance, the SAX-3/Robo proteins, involved in axon guidance, have a conserved domain architecture consisting of five immunoglobulin-like domains followed by three FN3-type domains [36,37]. There is evidence to suggest, however, that the domain structures of many extracellular proteins are more readily alterable than those of intracellular proteins. The saliva of the vampire bat *Desmodus rotundus* contains at least four homologues of human tissue-type plasminogen activator; however, each of these homologues contains fewer domains than their human counterpart (see Figure 1c). This situation is likely to have arisen as a result of accelerated rates of change of the domain architectures in order to acquire diverse physiological activities. Exceptionally fast protein evolution is believed to occur in gamete recognition proteins [38], including the modification of domain structures over very short time-spans. Gao and Garbers [39] report that equivalent sperm adhesion proteins in mouse and pig contain different numbers of MAM domains and von Willebrand D domains, although whether these genes are orthologous remains a moot point. A human polymorphism related to the number of kringle domains in apolipoprotein(a) is less in doubt [40].

As with intracellular signalling, in many cases, human extracellular proteins exhibit a greater variety of domain architectures when compared with *C. elegans*. For instance, the trypsin-like serine protease domain is combined with four other domains in *C. elegans*, whereas it can be found with 14 different domains in *H. sapiens*. As an example, human enterokinase is composed of six different domains, all of which are present in *C. elegans*, although none are found in the same protein acting as a serine protease domain (see Figure 1d).

Nuclear signalling

In order for cells to differentiate and thus perform specialised tasks in multicellular organisms, it is necessary to control the genes that are expressed in a particular cell. Thus, it is not surprising that families of transcription factors are greatly expanded in *C. elegans* relative to *S. cerevisiae* [41,42•].

It is evident that many transcription factor domains are unique or greatly expanded within a particular phylogenetic lineage. For instance, the SAND domain is found only in animal phyla [43], the RAV1 DNA-binding protein is currently found only in higher plants [44] and GAL4 DNA-binding domains are restricted to fungi, with large numbers present in yeast [45].

Histones are among the most highly conserved of proteins, yet, even here, differences can be found in multicellular organisms. The core histone, macroH2A, appears to have arisen from the combination of a H2A histone and an ancient domain that is conserved in bacteria. Phylogenetic analysis indicates that the two-domain fusion protein first appeared just prior to the branching of plants and animals [46]. Cloning of the *Drosophila* GCN5 histone acetylation transferase protein, which may play a role in regulating gene expression, has revealed the presence of a novel domain that is conserved among *Drosophila*, humans and mice [47]. This domain may interact with other transcription factors in the control of gene activation.

Evolution of function

The earliest multicellular organisms required both the preservation of functions that relate to unicellularity and the addition of new functions for both intracellular (e.g.

apoptosis) and extracellular (e.g. cell adhesion) compartmentalisations. The acquisition of novel functions whilst older functions are retained requires either *de novo* invention or the duplication of domains. Consequently, the genetically mobile domain might be considered the fundamental unit of function.

In many cases, it is difficult to detect which sequence families are homologous and, thus, from whence a new function was acquired. Many of the different domains mentioned in this review may form larger homologous groups, which, at present, are beyond our ability to detect using sequence-based methods. As more sequence data become available, it is likely that the number of 'new' domains that typify multicellularity will be reduced, as divergent sequence families are merged using newly available sequence information. For example, the laminin G, thrombospondin N and pentraxin families have recently been shown to comprise one large superfamily [48] and the WASP homology region 1 domain family has been shown to be part of a larger family, including Ran-binding domains [49].

The determination of structures can provide compelling evidence of homology. Although CARD, DED and DEATH domains lack interfamily sequence similarities, recent structure determinations and their functional roles as transducers of apoptotic signalling suggest that they are all homologous [50–53]. Thus, the apparent absence of these three domains from yeast might be better viewed as the absence of a single ancient homologue, rather than the absence of three separate families. Similarly, both the C1Q domain and the tumour necrosis factor (TNF) family of ligands appear to be absent in *C. elegans*, but present in humans; however, these domains have similar structures, in addition to sharing some short sequence motifs, and thus may be better viewed as being one superfamily [54]. Sequence analysis [55] and experimental results [56] have shown that the BTB/POZ domain (domain in broad-complex, tramtrack and bric-a-brac), found in zinc-finger-containing proteins, shares an evolutionary relationship with the potassium channel tetramerisation domain. The BTB/POZ domain is greatly expanded in *C. elegans* compared to the numbers present in *S. cerevisiae*. The analysis of Aravind and Koonin [55] strongly suggests that the BTB/POZ family has expanded independently in a number of phylogenetic lineages, after the divergence of yeast from other members of the eukaryotic crown group.

Once multidomain proteins had arisen and acquired useful functionality, they could be duplicated in their entireties and thus acquire new functions, either through different expression patterns or through changes of specificity. The Eph receptor kinases are implicated in various developmental processes and represent the largest known family of receptor tyrosine kinases (RTKs). The cytoplasmic part is composed of a tyrosine kinase domain and a C-terminal SAM domain [57,58]. The extracellular component consists of two FN3 domains and a N-terminal ligand-binding domain [59] (see Figure 1e). Whereas there is a single Eph receptor and two ephrins in *C. elegans* [60], 14 Eph receptor tyrosine kinases have been identified in humans, along with eight ephrin ligands. Such large-scale duplication events suggest that the process can occur independently of genome duplication.

Conclusions

Intracellular signalling modules are largely conserved between unicellular and multicellular life. In multicellular organisms, however, these domains are present both in larger numbers and in a wider variety of distinct contexts, reflecting an increased complexity of signalling pathways. Extracellular proteins, in contrast, contain many domains that have no detectable homologues in single-celled life. The newly available worm genome sequence hints that the diverse multicellular physiologies arise less from differences in extracellular domain repertoires and more both from lineage-specific expansions of such domain families and from the evolution of novel domain combinations.

Note added in proof

The findings cited in the text as (CP Ponting, L Aravind, J Schultz, P Bork, EV Koonin, unpublished data) have been recently accepted for publication [61].

Acknowledgements

RRC thanks Rob Russell (SmithKline Beecham) for helpful discussions. CPP is funded by a National Research Council (USA) Senior Associateship. This work is supported by the BMBF and the DFG.

References and recommended reading

Papers of particular interest, published within the annual period of review, have been highlighted as:

- • of special interest
- •• of outstanding interest

1. Koonin EV, Tatusov RL, Galperin MY: **Beyond complete genomes: from sequence to structure and function.** *Curr Opin Struct Biol* 1998, **8**:355-363.

2. Bork P, Dandekar T, Diaz-Lazcoz Y, Eisenhaber F, Huynen M, Yuan Y: **Predicting function: from genes to genomes and back.** *J Mol Biol* 1998, **283**:707-725.

3. •• The *C. elegans* sequencing consortium: **Genome sequence of the nematode *C. elegans*: a platform for investigating biology.** *Science* 1998, **282**:2012-2018.

The importance of this paper lies in the availability of the data that it describes – the first complete genome of a multicellular organism. For the time being, sequence analysis of all aspects of multicellularity will be reliant upon it. In addition, *C. elegans* represents one of the classic experimental models for developmental biology. Linking such experiments to sequences should facilitate the transfer of these data to other organisms.

4. Rubin GM: **The *Drosophila* genome project: a progress report.** *Trends Genet* 1998, **14**:340-343.

5. Pennisi E: **Fruit fly researchers sign pact with Celera.** *Science* 1999, **283**:767.

6. Meinke DW, Cherry JM, Dean C, Rounsley SD, Koorneef M: ***Arabidopsis thaliana*: a model plant for genome analysis.** *Science* 1998, **282**:671-682.

7. Huynen MA, Bork P: **Measuring genome evolution.** *Proc Natl Acad Sci USA* 1998, **95**:5849-5856.

8. Sidow A: **Gen(om)e duplications in the evolution of early vertebrates.** *Curr Opin Genet Dev* 1996, **6**:715-722.

9. Wittbrodt J, Meyer A, Schartl M: **More genes in fish?** *Bioessays* 1998, **20**:511-515.

10. Amores A, Force A, Yan Y, Joly L, Amemiya C, Fritz A, Ho RK, Langeland J, Prince V, Wang Y *et al.*: **Zebrafish *hox* clusters and vertebrate genome evolution.** *Science* 1998, **282**:1711-1714.

11. Cooke J, Nowak MA, Boerlijst M, Maynard-Smith J: **Evolutionary origins and maintenance of redundant gene expression during metazoan development.** *Trends Genet* 1997, **13**:360-364.

12. • Gibson TJ, Spring J: **Genetic redundancy in vertebrates: polyploidy and persistence of genes encoding multidomain proteins.** *Trends Genet* 1998, **14**:46-49.

Continuing the debate on paralogue persistence started in [11], the authors argue that some multidomain proteins may be inherently more resistant to decay by point mutation than single-domain proteins, owing to the existence of dominant negative phenotypes. See also the reply by Cooke following this article.

13. Mushegian AR, Garey JR, Martin J, Xiu LX: **Large-scale taxonomic profiling of eukaryotic model organisms: a comparison of orthologous proteins encoded by the human, fly, nematode, and yeast genomes.** *Genome Res* 1998, **8**:590-598.

14. • Chervitz SA, Aravind L, Sherlock G, Ball CA, Koonin EV, Dwight SS, Harris MA, Dolinski K, Mohr S, Smith T *et al.*: **Comparison of the complete protein sets of worm and yeast: orthology and divergence.** *Science* 1998, **292**:2022-2028.

This work is significant as it represents the first general comparison of the complete protein sets of *C. elegans* and *S. cerevisiae*. For certain types of proteins carrying out 'core biological processes', one-to-one orthologous relationships are identifiable, whereas for proteins involved in signal transduction and regulatory control, many novel domains are found. Although the results are unsurprising, they represent the starting point from which further analyses may proceed.

15. Maleszka R, De Couet HG, Gabor Miklos GL: **Data transferability from model organisms to human beings: insights from the functional genomics of the flightless region of *Drosophila*.** *Proc Natl Acad Sci USA* 1998, **95**:3731-3736.

16. • Schultz J, Milpetz F, Bork P, Ponting CP: **SMART, a simple modular architecture research tool: identification of signaling domains.** *Proc Natl Acad Sci USA* 1998, **95**:5857-5864.

The modular architectures of many of the more interesting proteins in multicellular organisms can be difficult to analyse using conventional database searching techniques. Systems such as SMART [19], ProfileScan [17•] and Pfam [18•] represent the easiest ways of analysing proteins at the domain level. The SMART system focuses on the sensitive detection of the intracellular and extracellular domains involved in signal transduction using manually curated alignments. Pfam covers all types of protein modules, at the expense of some sensitivity.

17. • Hofmann K, Bucher P, Falquet L, Bairoch A: **The PROSITE database, its status in 1999.** *Nucleic Acids Res* 1999, **27**:215-219.

See annotation to [16•].

18. • Bateman A, Birney E, Durbin R, Eddy SR, Finn RD, Sonnhammer ELL: **Pfam 3.1: 1313 multiple alignments and profile HMMs match the majority of proteins.** *Nucleic Acids Res* 1999, **27**:260-262.

See annotation to [16•].

19. Ponting CP, Schultz J, Milpetz F, Bork P: **SMART: identification and annotation of domains from signalling and extracellular protein sequences.** *Nucleic Acids Res* 1999, **27**:229-232.

20. Leonard CJ, Aravind L, Koonin EV: **Novel families of putative protein kinases in bacteria and archaea: evolution of the eukaryotic protein kinase superfamily.** *Genome Res* 1998, **8**:1038-1047.

21. Fraser A, James C: **Fermenting debate: do yeast undergo apoptosis?** *Trends Cell Biol* 1998, **8**:219-221.

22. Uren AG, Coulson EJ, Vaux D: **Conservation of baculovirus inhibitor of apoptosis repeat proteins (BIRPs) in viruses, nematodes, vertebrates and yeasts.** *Trends Biochem Sci* 1998, **5**:159-162.

23. • Aravind L, Dixit VM, Koonin EV: **The domains of death: evolution of the apoptosis machinery.** *Trends Biochem Sci* 1999, **24**:47-53.

This paper examines the phylogenetic distribution of protein domains found in the apoptotic machinery, in order to trace the evolutionary origins of this phenomenon. Apoptosis is an intriguing case study, involving many regulatory domains, some with apparently ancient origins, in a pathway that appears to have arisen relatively late in evolution.

24. Takayama S, Xie Z, Reed JC: **An evolutionarily conserved family of Hsp70/Hsc70 molecular chaperone regulators.** *J Biol Chem* 1999, **274**:781-786.

25. Jiang Y, Woronicz JD, Lin W, Goeddel DV: **Prevention of constitutive TNF receptor 1 signaling by silencer of death domains.** *Science* 1999, **283**:543-546.

26. Chen J, Rawlings ND, Stevens RAE, Barrett AJ: **Identification of the active site of legumain links it to caspases, clostripain and gingipains in a new clan of cysteine endopeptidases.** *FEBS Lett* 1998, **441**:361-365.

27. van der Biezen EA, Jones JDG: **The NB-ARC domain: a novel motif shared by plant resistance gene products and regulators of cell death in animals.** *Curr Biol* 1998, **8**:R226-R227.

28. del Pozo O, Lam E: **Caspases and programmed cell death in the hypersensitive response of plants to pathogens.** *Curr Biol* 1998, **8**:1129-1132.

29. Craven SE, Bredt DS: **PDZ proteins organize synaptic signaling pathways.** *Cell* 1998, **93**:495-498.

30. Bargmann CI: **Neurobiology of the *Caenorhabditis elegans* genome.** *Science* 1998, **282**:2028-2033.

31. Satterlee JS, Sussman MR: **Unusual membrane-associated protein kinases in higher plants.** *J Membr Biol* 1998, **164**:205-213.

32. Birkeland NK: **Cloning, molecular characterization, and expression of the genes encoding the lytic functions of lactococcal bacteriophage phi LC3: a dual lysis system of modular design.** *Can J Microbiol* 1994, **40**:658-665.

33. Nielson H, Engelbrecht J, Brunak S, von Heijne G: **Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites.** *Protein Eng* 1997, **10**:1-6.

34. Blaxter M: ***Caenorhabditis elegans* is a nematode.** *Science* 1998, **282**:2041-2046.

35. Tudor JE, Pallaghy PK, Pennington MW, Norton RS: **Solution structure of the ShK toxin, a novel potassium channel inhibitor from a sea anemone.** *Nat Struct Biol* 1998, **3**:317-320.

36. Kidd T, Brose K, Mitchell KJ, Fetter RD, Tessier-Lavigne M, Goodman CS, Tear G: **Roundabout controls axon crossing of the CNS midline and defines a novel subfamily of evolutionarily conserved guidance receptors.** *Cell* 1998, **92**:205-215.

37. Zallen JA, Yi BA, Bargmann CI: **The conserved immunoglobulin superfamily member SAX-3/Robo directs multiple aspects of axon guidance in *C. elegans*.** *Cell* 1998, **92**:217-227.

38. Vacquier VD: **Evolution of gamete recognition proteins.** *Science* 1998, **281**:1995-1998.

39. Gao Z, Garbers DL: **Species diversity in the structure of zonadhesin, a sperm-specific membrane protein containing multiple cell adhesion molecule-like domains.** *J Biol Chem* 1998, **273**:3415-3421.

40. Peynet J, Beaudeux JL, Woimant F, Flourie F, Giraudeaux V, Vicaut E, Launay JM: **Apolipoprotein(a) size polymorphism in young adults with ischemic stroke.** *Atherosclerosis* 1999, **142**:233-239.

41. Clarke ND, Berg JM. **Zinc fingers in *Caenorhabditis elegans*: finding families and probing pathways.** *Science* 1998, **282**:2018-2022.

42. • Ruvkun G, Hobert O: **The taxonomy of developmental control in *Caenorhabditis elegans*.** *Science* 1998, **282**:2033-2041.

Many early analyses of complete genomes have been little more than exercises in counting numbers of proteins and domains. This paper presents the beginnings of a more focused approach, concerning transcription factors and regulatory genes in specific developmental pathways, and a comparison with other organisms.

43. Gibson TJ, Ramu C, Gemund C, Aasland R: **The APECED polyglandular autoimmune syndrome protein, AIRE-1, contains the SAND domain and is probably a transcription factor.** *Trends Biochem Sci* 1998, **7**:242-244.

44. Kagaya K, Ohmiya K, Hattori OT: **RAV1, a novel DNA-binding protein, binds to bipartite recognition sequence through two distinct DNA-binding domains uniquely found in higher plants.** *Nucleic Acids Res* 1999, **27**:470-478.

45. Bork PJ, Gibson TJ: **Applying motif and profile searches.** *Methods Enzymol* 1996, **266**:162-184.

46. Pehrson JR, Fuji RN: **Evolutionary conservation of histone macroH2A subtypes and domains.** *Nucleic Acids Res* 1998, **26**:2837-2842.

47. Smith ER, Belote JM, Schiltz RL, Yang X, Moore PA, Berger SL, Nakatani Y, Allis CD: **Cloning of *Drosophila* GCN5: conserved features among metazoan GCN5 family members.** *Nucleic Acids Res* 1998, **26**:2948-2954.

48. Beckmann G, Hanke J, Bork P, Reich JG: **Merging extracellular domains: fold prediction for laminin G-like and amino-terminal thrombospondin-like modules based on homology to pentraxins.** *J Mol Biol* 1998, **275**:725-730.

49. Callebaut I, Cossart P, Dehoux P: **EVH1/WH1 domains of VASP and WASP proteins belong to a large family including Ran-binding domains of the RanBP1 family.** *FEBS Lett* 1998, **441**:181-185.

50. Huang B, Eberstadt M, Olejniczak ET, Meadows RP, Fesik SW: **NMR structure and mutagenesis of the Fas (APO-1/CD95) death domain.** *Nature* 1996, **384**:638-641.

51. Hofmann K, Bucher P: **The CARD domain: a new apoptotic signalling motif.** *Trends Biochem Sci* 1997, **5**:155-156.

52. Eberstadt M, Huang B, Chen Z, Meadows RP, Ng SC, Zheng L, Lenardo MJ, Fesik SW: **NMR structure and mutagenesis of the FADD (Mort1) death-effector domain.** *Nature* 1998, **392**:941-945.

53. Chou JJ, Masuo H, Duan H, Wagner G. **Solution structure of the RAIDD CARD and model for CARD/CARD interaction in caspase-2 and caspase-9 recruitment.** *Cell* 1998, **94**:171-180.

54. Shapiro L, Scherer PE: **The crystal structure of a complement-1q family protein suggests an evolutionary link to tumor necrosis factor.** *Curr Biol* 1998, **8**:335-338.

55. Aravind L, Koonin EV: **Fold prediction and evolutionary analysis of the POZ domain: structural and evolutionary relationship with the potassium channel tetramerization domain.** *J Mol Biol* 1999, **285**:1353-1361.

56. Ahmed KF, Engel CK, Prive GG: **Crystal structure of the BTB domain from PLZF.** *Proc Natl Acad Sci USA* 1998, **95**:12123-12128.

57. Stapleton D, Balan I, Pawson T, Sicheri F: **The crystal structure of an Eph receptor SAM domain reveals a mechanism for modular dimerization.** *Nat Struct Biol* 1999, **6**:44-49.

58. Thanos CD, Goodwill KE, Bowie JU: **Oligomeric structure of the human EphB2 receptor SAM domain.** *Science* 1999, **283**:833-836.

59. Himanen JP, Henkemeyer M, Nikolov DB: **Crystal structure of the ligand-binding domain of the receptor tyrosine kinase EphB2.** *Nature* 1998, **396**:486-491.

60. George SE, Simokat K, Hardin J, Chisholm AD: **The VAB-1 Eph receptor tyrosine kinase functions in neural and epithelial morphogenesis in *C. elegans*.** *Cell* 1998, **92**:633-643.

61. Ponting CP, Aravind L, Schultz J, Bork P, Koonin EV: **Eukaryotic signalling domain homologues in archaea and bacteria – ancient ancestry and horizontal gene transfer.** *J Mol Biol* 1999, in press.