Published in final edited form as:
Science. 1998 December 11; 282(5396): 2022–2028.

Comparison of the Complete Protein Sets of Worm and Yeast: Orthology and Divergence

Stephen A. Chervitz, **L. Aravind**, **Gavin Sherlock**, **Catherine A. Ball**, **Eugene V. Koonin**, **Selina S. Dwight**, **Midori A. Harris**, **Kara Dolinski**, **Scott Mohr**, **Temple Smith**, **Shuai Weng**, **J. Michael Cherry**, and **David Botstein**

S. A. Chervitz, G. Sherlock, C. A. Ball, S. S. Dwight, M. A. Harris, K. Dolinski, S. Weng, J. M. Cherry, and D. Botstein are in the Department of Genetics, Stanford University School of Medicine, Stanford, CA 94305–5120, USA. L. Aravind and E. V. Koonin are at the National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA. S. Mohr and T. Smith are in the Department of Biomedical Engineering, Boston University, Boston, MA 02115, USA

Abstract

Comparative analysis of predicted protein sequences encoded by the genomes of *Caenorhabditis elegans* and *Saccharomyces cerevisiae* suggests that most of the core biological functions are carried out by orthologous proteins (proteins of different species that can be traced back to a common ancestor) that occur in comparable numbers. The specialized processes of signal transduction and regulatory control that are unique to the multicellular worm appear to use novel proteins, many of which re-use conserved domains. Major expansion of the number of some of these domains seen in the worm may have contributed to the advent of multicellularity. The proteins conserved in yeast and worm are likely to have orthologs throughout eukaryotes; in contrast, the proteins unique to the worm may well define metazoans.

The nematode worm *Caenorhabditis elegans* is only the second eukaryote to have its genome completely sequenced (1). The first complete eukaryotic genome sequence, that of the budding yeast *Saccharomyces cerevisiae*, has been reported previously (2). Thus, for the first time, it is possible to compare the entire complements of encoded proteins of two highly diverged eukaryotic species, one of which is a unicellular microorganism and the other a multicellular animal.

The first result is quite surprising: Simple sequence comparisons allow one to predict, more often than not, orthologous pairs. In many cases, orthologous pairs can be confidently delineated even within families of highly similar proteins having many members from each organism. In fact, at the most stringent comparison value, ~57% of protein pairs contain just one worm and just one yeast protein. The set of highly conserved proteins is encoded by a minority of the open reading frames (ORFs) in each organism (~40% of yeast and 20% of worm; see Table 1). They carry out the core biological processes shared by these two eukaryotes, such as intermediary metabolism, DNA and RNA metabolism, protein folding, trafficking, and degradation.

The second result is more in line with expectation. Unlike yeast, the worm has a number of specialized, committed cell types with distinct and coordinated programs of gene expression. The differentiation of cell types in the animal is achieved through an elaborate developmental program that has been explored in detail in *C. elegans* (3). In contrast, yeast adapts dynamically to its environment by switching on different gene batteries in response to nutrient status, oxygen tension, mating pheromones, and other factors (4). It is widely believed that the physical basis of the developmental complexity of a multicellular

eukaryote is a system of protein regulators and signal transducers that is significantly more complex than that in unicellular organisms (5). Interspecies comparison of the protein domains used in regulation and signal transduction shows that although there is considerable sharing of domains, most of the proteins in which they appear are generally not orthologous. Increasing numbers of multidomain proteins during eukaryotic evolution are thought to have originated largely by domain shuffling (6). Indeed, we can predict evolutionary trends including (i) the evolution of new regulatory or signaling domains; (ii) evolution of new domain architectures from shared (presumably preexisting) domains; and (iii) expansion of particular domain families by a series of duplications.

The comparison of 6217 yeast ORFs with 19,099 worm ORFs produces much more information than can possibly be printed here. All of the underlying data, however, can be found in searchable form on our Web site within the Saccharomyces Genome Database (SGD) (genome-www.stanford.edu/Saccharomyces/worm/).

Shared Core Biology of Worm and Yeast: The Orthologs

We set out to compare and contrast the encoded protein complements by identifying both orthologous proteins (7), and shared and novel protein domains in yeast and worm. Distinguishing orthologs, which have evolved by vertical descent from a common ancestor and are presumed to carry out the same function (8), from paralogs, which arise by duplication and domain shuffling within a genome and hence may have divergent functions, is paramount when carrying out whole genome comparisons (9). Failure to do so can result in functional misclassification (10) and inaccurate molecular evolutionary reconstructions (11). In this part of our analysis, we did not attempt to detect distant homologs, which may be found by using less stringent criteria and more sensitive techniques (12).

We compared the predicted proteins of yeast and worm by first carrying out reciprocal WU-BLASTP (13) comparisons (that is, each predicted yeast protein against all the predicted proteins of the worm and vice versa). In every case in which a high-scoring pair (HSP) was detected, we collected all members of a group from both organisms by using several BLAST *P*-values as thresholds, as described in Table 1. The ORFs within each group were then ordered by similarity clustering with the CLUSTALW program (14) and displayed as multiple sequence alignments, rooted cluster dendrograms, and unrooted trees. Each of these displays for every comparison can be found on our Web site. The numbers of worm and yeast ORFs that fall into these clusters at various similarity thresholds are given in Table 1. Figure 1 graphically depicts the distribution of the sequences from the worm-yeast clusters within functional categories. The first significant (and some-what unexpected) observation is that the absolute number of ORFs for which we find worm and yeast homologs is about the same in each organism. At the highest level of similarity ($P < 10^{-100}$), approximately equal numbers of yeast and worm ORFs are present. This trend generally holds even at the lowest threshold we studied ($P < 10^{-10}$), where there are 2497 yeast ORFs (40% of total yeast ORFs) and 3653 worm ORFs (19% of total worm ORFs).

These observations suggest that the core biological processes of the two organisms are carried out by a similar number of proteins. It further suggests that the very large difference in the total number of different proteins encoded by the two organisms (~3.1-fold higher in worm) is not accounted for by endless close variations in the clusters found among the shared set, but instead are proteins that are substantially different in sequence [compare with (15)] and thus are likely to perform tasks that are specific to each organism. A subset of such organism-specific proteins, those associated with regulation and signal transduction, were investigated and found to support this idea [(16); and see below].

If many core biological processes of worm and yeast are indeed carried out by a comparable number of closely related proteins, then it might not be necessary to study the proteins (or the processes) in detail in both organisms. Instead, the annotation for the proteins involved in shared core biology (annotation that exists almost exclusively for yeast) might be transferable to the worm, provided that the orthologs between the two species are easily recognizable by sequence analysis alone. Functional conservation of proteins from different species was first demonstrated experimentally by showing that the mammalian RAS protein can substitute for yeast RAS in a RAS-deficient yeast strain (17). The worm RAS homolog *let-60* is involved in a variety of signaling processes (18) and is homologous to two yeast RAS genes (*RAS1* and *RAS2*), as described for many families below (see also the Web site). Although upstream regulators and downstream effectors of RAS may have diverged in the two organisms, it is likely that these orthologs may have a core biochemical function that is conserved, a prediction that can be easily tested in genetically tractable model organisms (19). In another example, yeast *CDC28* and worm *ncc-1* form an orthologous pair in the cyclin-dependent kinase family and have already been shown experimentally to be functionally interchangeable. When expressed in yeast, the protein encoded by *ncc-1* complements the G_2/M arrest of a *cdc28* temperature-sensitive mutation, illustrating functional conservation in vivo (20).

Table 1 shows that at each level of significance roughly half (611 of 1171 at $P < 10^{-10}$) of all the sequence similarity groups found by our reciprocal BLASTP procedure contain exactly two members. Because ascertainment of each group began with a yeast-worm HSP, these groups contain one worm and one yeast member. The availability of complete sequences for both worm and yeast makes it unlikely that we are missing large numbers of potential orthologs. It remains possible that the conservative similarity cutoffs used leave fast-evolving orthologs to be identified by more detailed analysis. Thus, most of the proteins contained in these 611 groups will turn out to be authentic orthologs, like the *CDC28/ncc-1* pair cited above.

Examination of the CLUSTALW output provides a comparably strong indication of many orthologous relationships within the remaining groups (560 of 1171 at $P < 10^{-10}$) that contain three or more members. From several hundred such families, six examples are illustrated in a rooted tree display (21,22) (Fig. 2). The first example (Fig. 2A) illustrates the two clusters of DNA-dependent RNA polymerases. In every case, the yeast and worm proteins form unambiguous pairs. In this instance, most of the cases for pairing are conclusive, because the RNA polymerase I and II subunits were independently identified in yeast and worm (23). In addition, the cluster [here done at $P < 10^{-20}$ (24)] contains the yeast polymerase III subunit paired with its presumed ortholog in the worm.

The second example (Fig. 2B) shows the cluster of DNA replication factor C subunits, which act as processivity factors for DNA polymerases δ and ε and load proliferating cell nuclear antigen (PCNA) onto DNA (25). This cluster has 12 members, and the pairing is entirely consistent with the idea that each member of each pair is orthologous to the other. The third example (Fig. 2C) shows a similar clustering of proteasome subunits (26). In this case there are 25 members of the cluster, which form 10 clear pairs, with three yeast and two worm sequences apparently unpaired. However, it seems probable that there is an additional orthology: yeast *PRE2* with the minimally diverged (recently duplicated?) worm sequences K05C4.1 and Y105E8A.jj. Accepting this, the 25 sequences yield 11 pairs.

The worm has 17 tubulin genes, compared to just 4 in yeast (Fig. 2D). Because the worm expresses specific tubulins for specific functions, a skewed worm:yeast ratio is to be expected. For instance, worm *tba-1* α-tubulin is selectively expressed in a set of mechanosensory and ventral-cord motor neurons during development. Conversely, yeast

express almost twice as many hexose transporters as worm, indicating the importance of sugar transport to *S. cerevisiae* (27). Both worm and yeast encode just one γ-tubulin, implying that whereas other tubulins may have become more specialized, γ-tubulin still functions only in a common core process.

The comparisons for actin and actin-like proteins give a quite different result (Fig. 2F). Although there are more classical actins in the worm than in yeast, several of the actin-related proteins (*ARP* genes) of yeast have what appear to be orthologs in the worm. Like γ-tubulin, they appear to carry out a core process shared by the two organisms. The true actins of the worm function in both muscular contraction and as cytoskeletal elements, so that the duplication and divergence of specialized actins was to be expected. Somewhat surprisingly, there is a yeast actin-related protein with no obvious counterpart in the worm, *ARP1* (28), which encodes a nuclear protein related to dynactin and centractin. This lack of orthology may be explained by the relatively unusual chromosome mechanics of *C. elegans*, whose chromosomes are holocentric and thus lack defined centromeres (29).

In the large cluster of HSP70 heat shock proteins (Fig. 2E), five subclusters can be recognized, each containing worm and yeast genes. The subclusters appear to reflect different localization or substrate specificities in yeast. One encodes yeast cytoplasmic HSP70 proteins (*SSA* genes); another encodes mitochondrial proteins (*SSC1*). A third encodes yeast cytoplasmic proteins that act on nascent peptides and associate with translating ribosomes (*SSB* genes) (30). Notably, the fourth group encodes genes that act as chaperones in the endoplasmic reticulum; Kar2p in yeast (31) and hsp-3 and hsp-4 in worm (32) have independently been characterized to have this function.

The nuclear-encoded mitochondrial proteins of worm and yeast provide a compelling example of orthologous pairs but also a remarkable case of the worm apparently missing orthologs for a set of important yeast proteins. Comparisons were performed with PSI-BLAST (33) and validated by demonstrating sequence similarity to *Escherichia coli* or *Methanoccus jannaschii* protein sequences. A total of 108 mitochondrial proteins from yeast have highly conserved homologs in worm (*P*-value scores $<10^{-39}$). These orthologous pairs can be assigned to diverse mitochondrial functions such as the TCA (tricarboxylic acid) cycle, electron transport, lipid metabolism, amino acid biosynthesis, intermediary metabolism, membrane transport, protein processing, RNA metabolism, and protein synthesis. Surprisingly, worm orthologs were identified for only 10 of the approximately 40 unique yeast mitochondrial ribosomal proteins (34). It seems possible that given the small size of mitochondrial ribosomal RNAs in the nematodes (35), the *C. elegans* mitochondrial ribosomes could contain a small number of proteins. However 10 proteins are unlikely to make a functional ribosome. It therefore remains to be determined whether more ribosomal protein genes are encoded in the worm genome but are missing in the currently defined gene complement, or if some have been displaced in the nematode mitochondrial ribosome by cytoplasmic ribosomal proteins.

Taken together, these observations show that for a substantial fraction of the yeast and worm genes, unequivocal, one-to-one orthologous relationships are readily identifiable. The simplest explanation for these results is that the proteins in this data set carry out core biological processes required by each organism. To test this idea, a functional classification for each of the proteins in this set was abstracted, mainly from the SGD (most of the yeast proteins in this set have some functional annotation) but also from the Web version of ACeDB (www.sanger.ac.uk/Projects/C_elegans). When this was done for the set of proteins at the level of $P < 10^{-50}$, 91% of the proteins could be classified. Of these, 79% could be assigned to rubrics fitting the description of core biological processes (Fig. 1). A more detailed scrutiny of orthologs in different functional categories indicates, however, that

certain central metabolic pathways (for example, those for the biosynthesis of several amino acids) that are present in yeast appear to be missing in the worm. This reflects the different nutritional requirements of the two organisms. Many of these functional designations are particularly reliable because they originate from experiments carried out directly with yeast.

Possibly the most important opportunity to emerge from these results is that annotation of protein functions and activities will be reliably transferable between organisms as disparate as yeast and worm by sequence analysis. With well-annotated genomes, the identification of orthologous pairs becomes a powerful analytical approach. Whereas biochemical and biological experiments must be done to unequivocally prove the functions of proteins, the wealth of data from sequence analyses allows researchers to better design experiments and avoid duplication of work done in other systems.

Distinguishing the Multicellular Worm from the Unicellular Yeast: The Divergence

The analysis thus far has concentrated on similarity of entire proteins. However, there are many instances in which domains, rather than entire proteins, are conserved. Yeast and worm have many core metabolic functions that are encoded in large multidomain proteins. A number of these are simple concatenations of various catalytic steps in the same pathway that, in the bacteria, is found on separate peptides. The identification of shared domains, as well as unique domain combinations, should provide important information on the functional divergences between these two organisms.

The primary interest in *C. elegans* is not in the shared core functions but in the functions characteristic of multicellularity. To investigate the worm proteins that are associated with such functions, we defined a set of 122 protein domains (36) that are widespread in eukaryotes and are associated with the regulation of gene expression and signal transduction (37). We then compared these domains in worm and yeast in terms of the number and domain architectures of the proteins in which they occur (38). The worm-yeast comparison highlights several distinct pathways for the evolution of innovations that seem to form the basis of complex signal transduction systems. In many cases, it appears that an ancestral regulator or an entire signaling system retains its general function but acquires new specificities after a series of duplications with subsequent divergence. We found instances of invention of domains de novo, recruitment of domains to novel forms of signal transduction, and amplification and diversification of domains already associated with signal transduction.

Relatively small but important sets of regulatory and signal transduction domains are found in *C. elegans* but not in *S. cerevisiae*, and vice versa (Table 2). These might well represent evolution of new regulatory or signal domains. The worm domains not found in yeast can generally be linked to the layers of complexity in signal transduction that accompany multicellularity. The most obvious examples are extracellular signaling and adhesion molecules, such as epidermal growth factor (EGF) and cadherin domains, first messengers such as FMRFamides and insulin-like peptides, and voltage- and ligand-gated channels such as degenerins. Other domains found in worm but not yeast include components of the programmed cell death machinery (such as the caspases). Also prominently figuring in this animal-specific class of regulatory domains are transcriptional regulators such as nuclear hormone receptors that are particularly numerous in the worm (Table 2). The genes encoding these classes of signaling domains appear to have evolved only in animals and have undergone varying degrees of duplication in even a simple multicellular organism such as *C. elegans*. Yeast encodes its own small set of fungal-specific regulatory domains (Table 2), of which the most prominent is the C6 finger, a DNA-binding domain.

The majority of the signaling domains, however, are detected in both yeast and worm. The method we used allowed us to amend the list of such conserved domains by discovering the yeast counterparts of several domains that have been thought to be unique for animals. Examples of important domains that were not previously detected in yeast include MATH, POZ, and arrestin (Table 2).

Several interesting examples illustrate how domains originally unrelated to signal transduction seem to have been recruited for important regulatory functions in animals. For example, the HINT (Hedgehog-INTein), PAIRED box, and POU domains appear to have been derived from selfish elements or transposons. The HINT domain (39) is found in single copy in yeast where it appears to be a selfish genetic element (intein) in *TFP1*, a vacuolar ATPase (adenosine triphosphatase) subunit gene. By contrast, the worm has 11 copies, always as a part of a molecule that is probably autocatalytically cleaved to produce an extracellular regulator (40). The further history of this domain includes the origin of Hedgehog, a key regulator of positional information in vertebrate and insect development. The history of PAIRED box and POU domains (41), which are missing in yeast but prominent in the predicted worm transcription regulators, appears similar to HINT. These DNA-binding domains are specifically related to the helix-turn-helix domains of the transposases of animal and bacterial transposons (42), which probably indicates the route whereby they invaded the ancestral animal genome.

Examples of regulatory domains that are detectably conserved in *C. elegans* and *S. cerevisiae*, but nevertheless act in worm signal transduction pathways not found in yeast, include the immunoglobulin, FN3, LRR, and vWA domains. In yeast, these domains act within the cell, in DNA binding or intracellular protein–protein interactions, whereas in the worm they become prominent extracellular adhesion and signaling modules (Table 2). The SH2 domain may have some conserved functions in yeast and in the worm, as indicated by the conservation of the entire domain architecture of the SH2-containing transcription factor Spt6p. However, the best known role of SH2, in the tyrosine phosphorylation signaling system, is clearly an innovation for which this domain had been recruited only in animals.

A numerical comparison shows that in many cases, the number of proteins with the given domain in worm and yeast is about the same when normalized by the total gene numbers (Fig. 3). Against this background, the marked expansion of several domains in the worm is striking (Fig. 3 and Table 2). The domains with a disproportionate excess in the worm include (i) a small, distinct set of protein-protein interaction domains, such as the intracellular domains (MATH, POZ, PDZ, and LIM), and the largely extracellular domains (FN3, LRR, and vWA); (ii) the phosphotyrosine signaling system–tyrosine kinases, phosphotyrosine phosphatases, SH2, and PTB (two types of phosphotyrosine-binding domains; the PTB domain was not detected in yeast); (iii) the cyclic nucleotide monophosphate (cNMP)–dependent signaling system—the cNMP cyclases, phosphodiesterases, and cNMP-binding domains; (iv) homeodomains; (v) calmodulin-type EF-hand domains; (vi) potassium channels; and (vii) 7TM receptors. Notably, the MATH and POZ domains showed the quantitatively greatest expansion in the worm (Table 2 and Fig. 3). The entire range of the functions of these domains remains to be clarified, but a large family of worm proteins that contain these two domains combined are likely to be involved in specific aspects of chromosome organization (43). Within the chosen set of domains, there are no significant family extensions in yeast compared to *C. elegans* (Table 2 and Fig. 3). In certain cases, however, such as the C_2H_2 Zn finger domain–containing proteins, the numerous *C. elegans* and *S. cerevisiae* proteins form separate intraspecies groups, suggesting independent duplication histories.

Conclusions

This first reciprocal analysis of two complete eukaryotic genome sequences has produced two kinds of results. First, it is clear that a comparable number of orthologous proteins carry out the core functions of both *S. cerevisiae*, a unicellular free-living budding yeast, and *C. elegans*, a multicellular nematode. Second, most of the signaling and regulatory genes known or expected to be involved in multicellularity have no yeast orthologs, even though they may contain domain sequences shared with yeast. Thus, virtually all biological processes characteristic of multicellular life are performed by proteins that are not close variants of proteins responsible for the core processes, even though they might share some domains.

Both of these conclusions depend strongly on having virtually complete sequences for both organisms. If only a fraction of the total sequence is known, there is no way to make inferences concerning failure to find a homolog. Likewise, if a comparable domain arrangement is not detected in an incomplete sequence, it is impossible to conclude that this domain arrangement is absent.

These findings have clear inferences with respect to the gene complement of the common ancestor of animals and fungi. Clearly, this hypothetical ancestor encoded all the conserved proteins responsible for core functions. Most known signaling and regulatory domains must have been already encoded in this ancestral genome. Furthermore, certain important, versatile regulators, such as AAA superfamily ATPases and SWI/SNF2 helicases, are conserved in yeast and worm in terms of their absolute rather than normalized counts. The respective sets of proteins appear to show one-to-one orthologous relationships, which suggests that their functions have been established already in the common ancestor of fungi and animals. At the level of entire signal transduction pathways, however, there is relatively little conservation between worm and yeast. Taken together, these observations suggest that the common ancestor possessed signal transduction systems that were distinct from those seen either in yeast or in the worm, although they might have resembled the less elaborate yeast pathways more closely. A notable observation is the small number of conserved transcription factors in yeast and worm, which suggests that the common ancestor encoded only a small fraction of the extant transcription regulators. There may be two equally interesting explanations for this: (i) the common ancestor had only a rudimentary system for transcription regulation and (ii) the ancient regulators have been displaced by new ones that evolved after the radiation of the major eukaryotic lineages.

Finally, the basic assumption that the so-called “model organisms” will provide reliable functional annotation for the human DNA sequence is strongly supported by our observations. First, the sum of the biology of worm and yeast can be obtained efficiently by studying core functions largely in yeast and signal transduction largely in the worm, with virtually no overlap. Second, the evolutionary distance (and biological diversity) between yeast and worm did not interfere with the finding of orthologs and shared domains, making it likely that a robust chain of annotation is possible through all of the eukaryotes.

Acknowledgments

We thank J. Hodgkin, R. Horvitz, J. Kimble, and the editors of *Science* for the invitation to write this paper, D. Lipman for suggesting the collaborations, and K. Anders for helpful discussions. We are especially grateful to R. Durbin (Sanger Centre) and L. Hillier (Genome Sequencing Center) for providing sequence information and for their cooperation. The SGD is supported by a P41 national resources grant HG01315, from the National Human Genome Research Institute at the U.S. NIH. S.A.C. is supported by training grant PHS HG 00044. T.S. is supported by grant DE-FG02-98ER62558 from the Department of the Environment.

References and Notes

1. *C. elegans* Sequencing Consortium. Science. 1998; 282:2012. [PubMed: 9851916]
2. Goffeau A, et al. ibid. 1996; 274:546.
3. Horvitz HR, Sulston JE. Genetics. 1980; 96:435. [PubMed: 7262539] Sulston JE, White JG. Dev. Biol. 1980; 78:577. [PubMed: 7190941] Kenyon C. Science. 1988; 240:1448. [PubMed: 3287621] Sternberg PW. Adv. Genet. 1990; 27:63. [PubMed: 1971988] Sternberg PW, Felix MA. Curr. Opin. Genet. Dev. 1997; 7:543. [PubMed: 9309188]
4. Nasmyth K, Shore D. Science. 1987; 237:1162. [PubMed: 3306917] Herskowitz I. Microbiol. Rev. 1988; 52:536. [PubMed: 3070323] Marsh L, Herskowitz I. Cold Spring Harbor Symp. Quant. Biol. 1988; 53:557. [PubMed: 3151177] Marsh L, Neiman AM, Herskowitz I. Annu. Rev. Cell Biol. 1991; 7:699. [PubMed: 1667085] Nasmyth K. Trends Genet. 1996; 12:405. [PubMed: 8909137]
5. Gerhart, J.; Kirschner, M. Cells, Embryos, and Evolution. Malden, MA: Blackwell; 1997.
6. Doolittle RF. Annu. Rev. Biochem. 1995; 64:287. [PubMed: 7574483]
7. Orthology is not necessarily a one-to-one relationship; a unique gene in one species may be the ortholog of a gene family in another species. The issue is further confounded by the fact that many proteins, particularly in eukaryotes, contain multiple domains that have a degree of evolutionary independence and are found in different combinations. Thus, the detection of even very high similarity between protein sequences from two species does not guarantee that the proteins in question are genuine orthologs with a conserved domain architecture (10). To diminish (but not eliminate) the latter problem, we required that the members of each protein pair be aligned through at least 80% of their lengths.
8. Fitch WM. Syst. Zool. 1970; 19:99. [PubMed: 5449325]
9. Koonin EV, Mushegian AR, Galperin MY, Walker DR. Mol. Microbiol. 1997; 25:619. [PubMed: 9379893]
10. Hennikof S, et al. Science. 1997; 278:609. [PubMed: 9381171] Tatusov RL, Koonin EV, Lipman DJ. ibid. :631.Galperin MY, Koonin EV. Silico Biol. 1998; 1:7. www.bioinfo.de/isb/1998/01/0007.
11. Doolittle RF, et al. Science. 1996; 271:470. [PubMed: 8560259] Feng DF, Cho G, Doolittle RF. Proc. Natl. Acad. Sci. U.S.A. 1997; 94:13028. [PubMed: 9371794]
12. The data set used for these comparisons was the 16 October 1998 worm protein data set from the Sanger Centre and the ORF translations in the 28 October version of the SGD (both are available from the *Science* Web site as well as SGD). Because the prediction of *C. elegans* protein sequences had, on 16 October, yet to be corrected by rigorous experimental analysis, our reliance on these predictions may result in the loss of some subset of *C. elegans* proteins. However, using the subset of yeast proteins for which we had identified no worm homolog, we performed BLAST searches against six frame translations of the entire worm DNA sequence (finished sequence from the Sanger Centre Web site as of 3 November 1998) and identified no additional homologs at the $P < 10^{-10}$ level with the >80% alignment requirement (J. M. Cherry, unpublished data). Supplemental information regarding the analysis is available at www.sciencemag.org/feature/data/c-elegans.shl for a general overview and at www.sciencemag.org/feature/data/985134.shl for information specific to this review.
13. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. J. Mol. Biol. 1990; 215:403. [PubMed: 2231712] Gish W, States DJ. Nature Genet. 1993; 3:266. [PubMed: 8485583] Karlin S, Altschul SF. Proc. Natl. Acad. Sci. U.S.A. 1993; 90:5873. [PubMed: 8390686] Altschul SF, Gish W. Methods Enzymol. 1996; 266:460. [PubMed: 8743700] Version 2.0a19MP-WashU of BLAST was used, with the XNU and SEG filters, BLOSUM62 scoring matrix, with gapping on, and other parameters set to default values.
14. Thompson JD, Higgins DG, Gibson TJ. Nucleic Acids Res. 1994; 22:4673. [PubMed: 7984417] Higgins DG, Thompson JD, Gibson TJ. Methods Enzymol. 1996; 266:383. [PubMed: 8743695] Version 1.74 of CLUSTALW was used with the BLOSUM substitution matrices.
15. Green P, et al. Science. 1993; 259:1711. [PubMed: 8456298]
16. Clarke ND, Berg JM. ibid. 1998; 282:2018.
17. Kataoka T, et al. Cell. 1985; 40:19. [PubMed: 2981628]

18. Chamberlin HM, Sternberg PW. Development. 1994; 120:2713. [PubMed: 7607066] Church DL, Guan KL, Lambie EJ. ibid. 1995; 121:2525.Gutch MJ, Flint AJ, Keller J, Tonks NK, Hengartner MO. Genes Dev. 1998; 12:571. [PubMed: 9472025] Sundaram M, Yochem J, Han M. Development. 1996; 122:2823. [PubMed: 8787756] Yochem J, Sundaram M, Han M. Mol. Cell. Biol. 1997; 17:2716. [PubMed: 9111342]
19. Botstein D, Chervitz SA, Cherry JM. Science. 1997; 277:1259. [PubMed: 9297238] Botstein D, Fink GR. ibid. 1988; 240:1439.
20. Mori H, Palmer RE, Sternberg PW. Mol. Gen. Genet. 1994; 245:781. [PubMed: 7830726]
21. Although there are strong theoretical reasons for preferring the unrooted tree, we show the rooted trees because they are easier to display compactly and more clearly represent the relationships at the tips of the branches, where the assessment of orthology is made. These are, in fact, just representations of unrooted trees with rooting that should be considered arbitrary.
22. Felsenstein J. Methods Enzymol. 1996; 266:418. [PubMed: 8743697]
23. Rogalski TM, Riddle DL. Genetics. 1988; 118:61. [PubMed: 8608933] Archambault J, Friesen JD. Microbiol. Rev. 1993; 57:703. [PubMed: 8246845]
24. Figure 2A shows the CLUSTALW alignment at $P < 10^{-20}$ because at the $P < 10^{-10}$ the yeast protein Spt5p is included, paired with a presumed worm ortholog even though the similarity of these to RNA polymerases is, upon further study, clearly spurious. This artifact, due to low complexity in the Spt5p amino acid sequence, is avoidable by more aggressive filtering, applying the >80% alignment requirement as well as increasing the stringency; each of these measures exacts a cost in information as well. It illustrates that any alignment result has to be studied for robustness with regard to both stringency and filtering.
25. Mossi R, Hubscher U. Eur. J. Biochem. 1998; 254:209. [PubMed: 9660172]
26. Tanaka K. Biochem. Biophys. Res. Commun. 1998; 247:537. [PubMed: 9647729]
27. Chervitz SA, et al. data not shown.
28. Clark SW, Meyer DI. Nature. 1992; 359:246. *J. Cell Biol*. **127**,129 (1994). [PubMed: 1356230]
29. Herman RK, Cari CK, Hartman PS. Genetics. 1982; 102:379. [PubMed: 6890921]
30. Nelson RJ, Ziegelhoffer T, Nicolet C, Werner-Washburne M, Craig EA. Cell. 1992; 71:97. [PubMed: 1394434]
31. Normington K, Kohno K, Kozutsumi Y, Gething MJ, Sambrook J. ibid. 1989; 57:1223.
32. Heschl MFP, Baillie DL. Comp. Biochem. Physiol. 1990; 96:633.
33. Altschul SF, et al. Nucleic Acids Res. 1997; 25:3389. [PubMed: 9254694]
34. The mitochondrial ribosomal protein orthologs have been missed by the automatic comparison procedure primarily because they contain nonconserved NH_2-terminal import peptides as well as COOH-terminal tails (28).
35. Okimoto R, Macfarlane JL, Wolstenholme DR. J. Mol. Evol. 1994; 39:598. [PubMed: 7528811]
36. The domains were primarily from the SMART database [J. Schultz, F. Milpetz, P. Bork, C. P. Ponting, *Proc. Natl. Acad. Sci. U.S.A*. **95**, 5857 (1998)], to which several domains were added. We have not attempted to cite the literature for each domain, but refer the reader to the SMART database. See also www.bork.embl-heidelberg.de/Modules_db/special_annotation_page.html
37. Bork P, Schultz J, Ponting CP. Trends Biochem Sci. 1997; 22:296. [PubMed: 9270302]
38. To obtain robust counts for each of the domains in the yeast and worm protein sets, we compared representative sequences of each domain to the nonredundant protein database (National Center for Biotechnology Information, National Institutes of Health, Bethesda, MD) using the PSI-BLAST program, and the resulting position-dependent weight matrices (profiles) were saved. The number of search iterations and the cutoff for inclusion of sequences in the profile were adjusted individually for each domain. For most widespread domains, several profiles were constructed to ensure complete coverage. The profiles were then compared separately to the yeast and worm protein databases. Typically, the random expectation value of 0.01 was used as the criterion for domain identification, but the search results were additionally scrutinized for the conservation of patterns typical of the respective domain, to ensure the elimination of any false positives. The profiles for each of the domains are available at the Web site. They can be obtained by FTP and used for PSI-BLAST searches.
39. Hall TM, et al. Cell. 1996; 91:85. [PubMed: 9335337]

40. Porter JA, et al. ibid. 1996; 86:21.

41. Dahl E, Koseki H, Balling R. Bioessays. 1997; 19:755. [PubMed: 9297966] Ryan AK, Rosenfeld MG. Genes Dev. 1997; 11:1207. [PubMed: 9171367]

42. Franz G, Loukeris TG, Dialektaki G, Thompson CR, Savakis C. Proc. Natl. Acad. Sci. U.S.A. 1994; 91:4746. [PubMed: 8197129]

43. Nagai Y, et al. FEBS Lett. 1997; 418:23. [PubMed: 9414087] Aravind L, Koonin EV. J. Mol. Biol. in press.

44. Cherry JM. Nature. 1997; 387:67. [PubMed: 9169866]

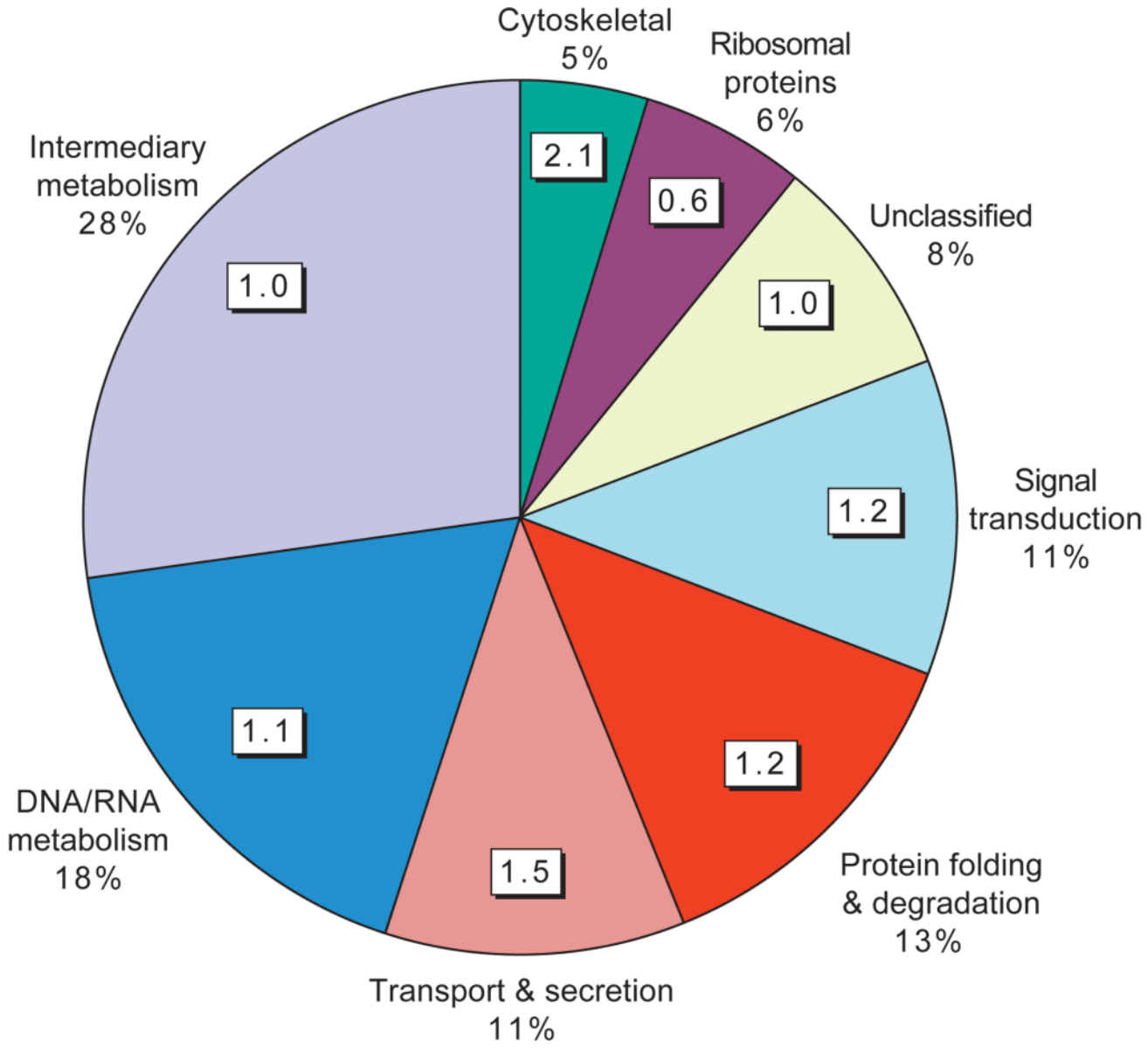

Fig. 1.

Distribution of core biological functions conserved in both yeast and worm. Yeast and worm protein sequences were clustered into closely related groups (BLASTP $P < 1 \times 10^{-50}$, with the >80% aligned length constraint) as described in the legend to Table 1. Each sequence group (including groups with two or more sequences) was assigned into a single functional category, relying primarily on the functional annotations for the yeast genes in SGD when available (44). The unclassified category contains groups of sequences without annotation. The boxed number within each category reflects the ratio of worm to yeast proteins for that category.

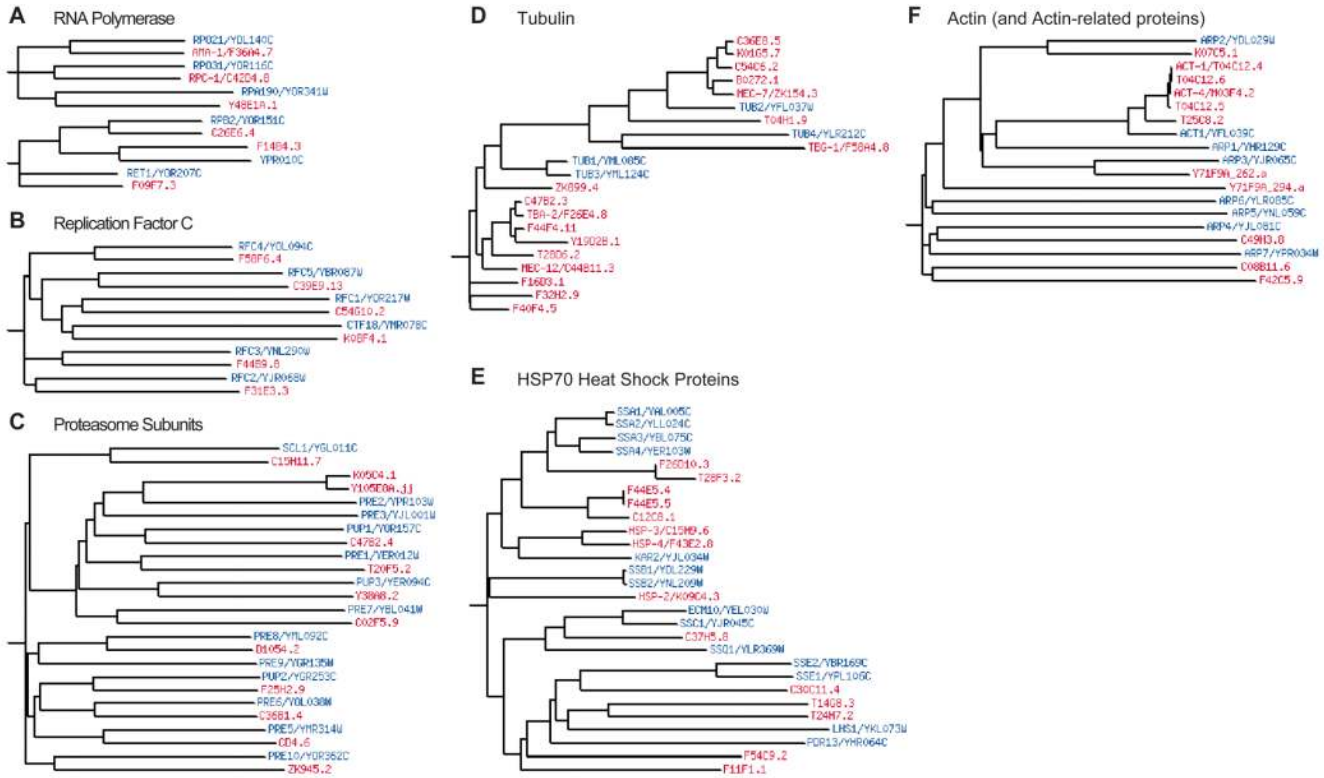

Fig. 2.

Orthologous core biological functions in yeast and worm. Representative sequence groups are shown as rooted CLUSTALW Neighbor-Joining trees, clustered as described in the legend to Table 1, at a similarity level indicated after each description. Gene names are color-coded (blue, *S. cerevisiae*; red, *C. elegans*). (**A**) RNA polymerase. (**B**) Replication factor C. (**C**) Proteasome subunits. (**D**) Tubulin. (**E**) HSP70 heat shock proteins. (**F**) Actin and actin-related proteins. Hyperlinked versions of these figures are available at the SGD Web site described in the text. Trees were created by means of CLUSTALW (14) with default parameters, which use the BLOSUM series of weight matrices. Trees were drawn with a combination of the Phylip (22) and gd (Boutell.Com, www.boutell.com) software packages. This table gives only examples from the table on the Web site.

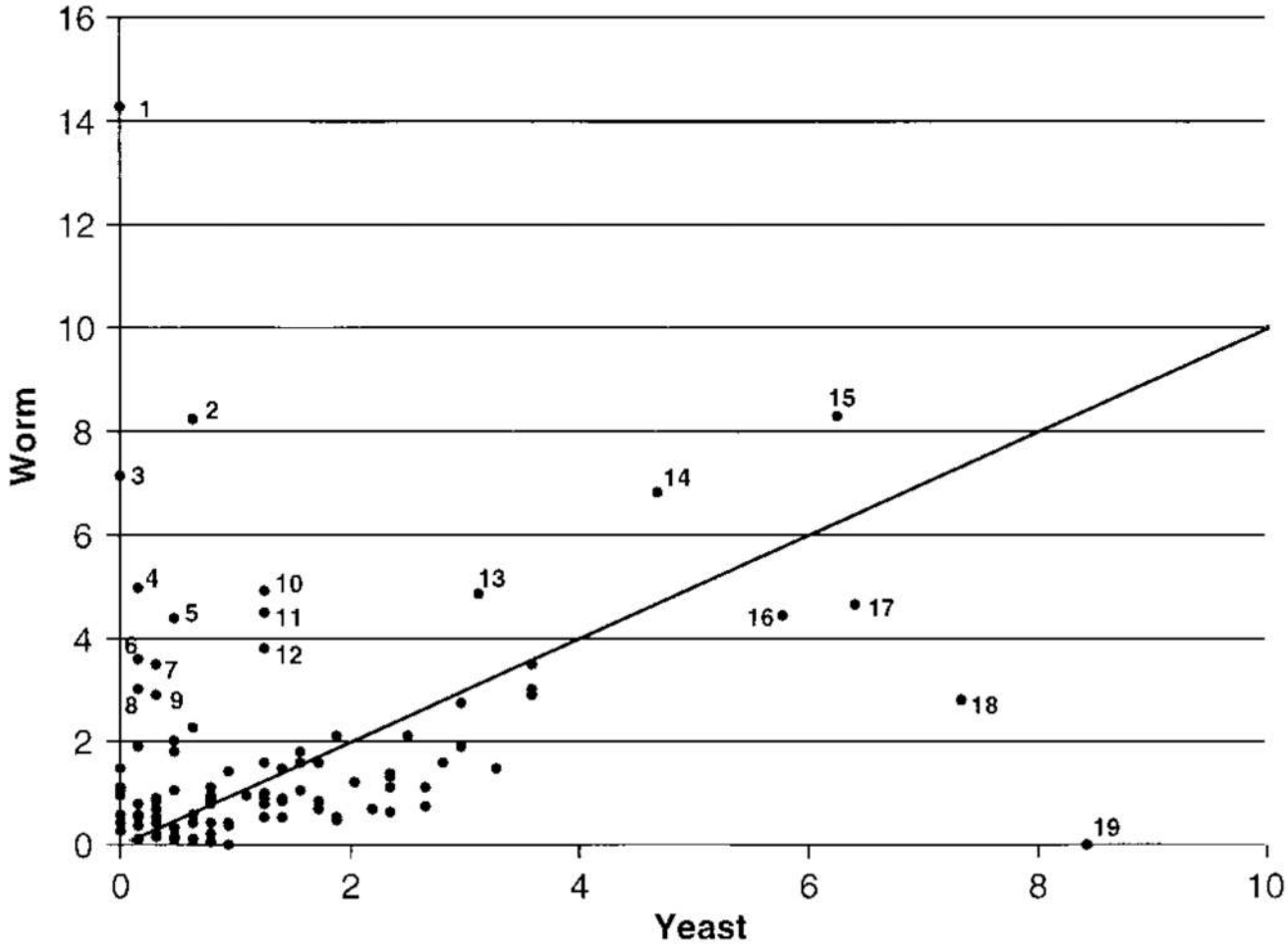

Fig. 3.

Normalized counts of the common regulatory and signal transduction domains in yeast and the nematode *C. elegans*. The data are from the complete version of Table 2 that is available on the Web site. The axes show the number of proteins with the given domain per 1000 genes. 1, POZ domains; 2, EGF; 3, MATH domain; 4, phosphotyrosine phosphatase; 5, homeodomains; 6, leucine-rich repeats; 7, calmodulin; 8, PDZ domains; 9, voltage-gated channels; 10, ankyrin repeats; 11, RING domain; 12, C6 fingers; 13, nuclear hormone receptors; 14, AAA-type ATPases. The two domains that are most abundant in yeast, namely serine-threonine protein kinases and WD40 domains (Table 2), were not used for this plot, in order to improve resolution for the other points.

Table 1

Conservation of yeast and worm protein sequences. Two reciprocal BLASTP analyses were performed, first with each of the 6217 yeast ORFs as a query against the worm ORF dataset (19,099 ORFs) (yeast versus worm), and second with each worm ORF as a query against the yeast ORF data set (worm versus yeast). For each yeast query sequence with a significant BLASTP worm hit (based on a conservative *P*-value cutoff, see below) in the yeast-versus-worm analysis, a sequence group was formed by combining the yeast query with all of its worm hits. This list was augmented by adding the yeast hits produced by each of these worm hits when used as a query in the worm-versus-yeast BLASTP analysis, imposing the same *P*-value cutoff. Analogous groups were constructed by starting with each worm query sequence from the worm-versus-yeast analysis, again with the same *P*-value cutoff. All of the above sequence groups were processed together, removing redundant sequences within each group and coalescing different groups if they contained any common sequence or sequences. Within all groups collected for a given *P*-value data set, each yeast and worm sequence will occur only once. Different maximum *P*-value cutoffs were set for the initial collection of hits (1×10^{-10}, 1×10^{-20}, 1×10^{-50}, and 1×10^{-100}). Sequence groups were also constructed with the additional constraint that 80% or more of each query sequence be aligned; results were similar to those without the aligned length constraint and can be viewed on our Web site.

P-value	Sequence groups		Yeast ORFs (%) (*n* = 6217)	Worm ORFs (%) (*n* = 19099)
	Total	>2 members		
1×10^{-100}	236	79	330 (5.3)	370 (1.9)
1×10^{-50}	552	230	888 (14.3)	1094 (5.7)
1×10^{-20}	984	442	1848 (30.0)	2479 (13.0)
1×10^{-10}	1171	560	2497 (40.0)	3653 (19.0)

Table 2

Unique and conserved regulatory and signal transduction domains in yeast and worm.

Domain	Brief description	Y*	W†	R‡
Domains found only in the worm				
PTB	Phosphotyrosine binding domain	0	11	-
Nuclear hormone receptors (NHR)	Transcription factors with ligand and DNA binding Zn-finger domains	0	270	-
EGF	Calcium-binding cysteine-rich repeats seen in epidermal growth factor and numerous other extracellular proteins	0	135	-
Degenerins	Amiloride-sensitive NA^+ channels	0	28	-
T-box	DNA-binding domain of transcription factors	0	21	-
FMRFamides	Neuropeptides	0	20	-
Cadherin	Calcium-dependent cell adhesion module	0	18	-
Paired box	DNA-binding domain with 2 helix-turn-helix (HTH) units	0	18	-
SMAD	Transcription factors	0	8	-
Insulin-like peptides	Peptide hormones	0	7	-
Laminin NT	N-terminal globular domain of the extracellular matrix protein laminin	0	5	-
Domains found only in yeast				
APSES	A fungal-specific DNA-binding domain seen in Swi4p	6	0	-
C6	A fungal-specific binuclear Zn-binding cluster	54	0	-
Conserved domains				
MATH	Globular domain shared by Meprin (metalloproteases) and the TRAFs (apoptosis effectors)	1	94	31.8
Voltage gated Channels	Ion channels typified by K^+ channel shaker	1	68	23.0
SH2 domain	Phosphotyrosine-binding domain (Src homology domain 2)	1	57	19.3
POZ	Protein-protein interaction domain first identified in Zn finger transcription regulators (e.g. Tramtrack) and poxvirus proteins also containing the KELCH repeats	4	156	13.2
cNMp cyclase	Catalytic domain of cyclic nucleotide (cAMP and cGMP) biosynthesis enzymes	1	36	12.2
Conserved domains				
PDZ domain	Protein-protein interaction domain binding C-termini of membrane-associated polypeptides	2	66	11.2
PTPase	Phosphotyrosine phosphatase	3	83	9.4
FNIII	Fibronectin III domain, adhesion module in animal extracellular proteins	2	55	9.3
LON protease	Serine protease domain frequently combined with a AAA ATPases	1	15	5.1
LIM domain	Cysteine-rich domain involved in protein-protein interactions and possibly DNA binding	3	38	4.3
Homeodomain	HTH-containing DNA-binding domain	8	93	3.9
SCP domain	Cysteine-rich module seen in snake/insect toxins and plant pathogenesis response protein	3	34	3.8
HINT domain	Domain shared by hedgehog and inteins; involved in autoproteolysis and protein splicing	1	11	3.7
vWA domain	Von Willebrand factor A domain; Mg^{2+}-binding adhesion module	4	43	3.6
LRR	Leucine-rich repeat involved in protein-protein interaction	8	85	3.6

Domain	Brief description	Y*	W†	R‡
Domains found only in the worm				
ZZ	Cysteine-rich module seen in transcriptional adaptors like ADA2P	1	10	3.4
Calmodulin-like EF hands	Calcium-binding helical acidic domains	8	72	3.0
cNMP-binding domain	Cyclic nucleotide-binding domain; the protein kinase A regulatory subunit	2	17	2.9
C2H2 finger	Zn-chelating DNA binding domain (classic Zn finger)	40	157	1.3
Protein kinase (STY)	Catalytic domain of protein kinases phosphorylating serine, threonine and tyrosine	118	435	1.2
AAA ATPases	A superfamily of ATPases including regulators of replication and ATP-dependent chaperones	47	53	0.4
WD40 repeats	β-propeller-forming repeat motif with a typical WD signature	110	127	0.4
SWI/SNF helicase	Large ATPases (member of the helicase supefamily II) involved in chromatin dynamics and repair	17	21	0.4

*Number of proteins containing the given domain in yeast.

†Number of proteins containing the given domain in the worm.

‡Ratio of the number of proteins with the given domain in the worm to the number in yeast, normalized by the total number of genes: $R = W*6.2/Y*18.9$.