GENOMICS **46,** 200–216 (1997)
ARTICLE NO. GE974989

Analysis of Protein Domain Families in *Caenorhabditis elegans*

Erik L. L. Sonnhammer[1,2] and Richard Durbin

Sanger Centre, Hinxton Hall, Cambridge CB10 1SA, United Kingdom

Received April 18, 1997; accepted August 26, 1997

The *Caenorhabditis elegans* genome sequencing project has completed over half of this nematode's 100-Mb genome. Proteins predicted in the finished sequence have been compiled and released in the database Wormpep. Presented here is a comprehensive analysis of protein domain families in Wormpep 11, which comprises 7299 proteins. The relative abundance of common protein domain families was counted by comparing all Wormpep proteins to the Pfam collection of protein families, which is based on recognition by hidden Markov models. This analysis also identified a number of previously unannotated domains. To investigate new apparently nematode-specific protein families, Wormpep was clustered into domain families on the basis of sequence similarity using the Domainer program. The largest clusters that lacked clear homology to proteins outside Nematoda were analyzed in further detail, after which some could be assigned a putative function. We compared all proteins in Wormpep 11 to proteins in the human, *Saccharomyces cerevisiae*, and *Haemophilus influenzae* genomes. Among the results are the estimation that over two-thirds of the currently known human proteins are likely to have a homologue in the whole *C. elegans* genome and that a significant number of proteins are well conserved between *C. elegans* and *H. influenzae*, that are not found in *S. cerevisiae*. © 1997 Academic Press

INTRODUCTION

Genome sequencing projects produce data that open up many new areas of investigation, such as the analysis of all the proteins encoded in a genome. Knowing the complete set of proteins in an organism is important for studies of protein evolution and function and for conclusive comparisons of the proteins present in different organisms (Koonin *et al.*, 1996b; Tatusov *et al.*, 1996). In this paper, we take a closer look at the proteins encoded in the genome of a higher eukaryote, the nematode *Caenorhabditis elegans* (Wilson *et al.*, 1994; Hodgkin *et al.*, 1995; Waterston and Sulston, 1995), by systematic functional classification, clustering of gene families, and comparison to other genomes.

Functional classification based on the protein sequence alone exploits the preexisting annotation of homologous protein sequences with a known function. Searching single-sequence databases with pairwise methods is by far the most common technique. In many cases, however, pairwise methods are not as sensitive as multiple alignment methods (Gribskov *et al.*, 1987; Henikoff and Henikoff, 1991; Krogh *et al.*, 1994; Eddy, 1996). Furthermore, if the annotation is extracted automatically from matches to single sequences, it may be misleading or inappropriate, for instance if the annotation pertains to a different domain or is not functionally relevant. A different approach, which is not as comprehensive but is sometimes more sensitive and generally less ambiguous regarding the extent of homologous domains, is to search a database of preassembled multiple alignments of protein domain families. An example of such a database is Pfam (Sonnhammer *et al.*, 1997), which is based on recognition by hidden Markov models (HMMs) (Krogh *et al.*, 1994). Matches to Pfam families provide a more general level of annotation. Here we describe the Pfam-based classification of the proteins predicted so far (about 50%) in the *C. elegans* genome.

A fundamental principle of protein evolution is that new protein functions can arise by the duplication of a gene and subsequent specialization of the "daughters." In most cases the detailed functions, or roles, of the daughters are different, although their structure and catalytic mechanisms are analogous. In fact, the majority of proteins in higher eukaryotic genomes, and 30–50% in prokaryotes (Brenner *et al.*, 1995; Koonin *et al.*, 1996b), have clearly recognizable "siblings" that are products of gene duplication. Such homologues are called paralogues, while proteins in different organisms that diverged due to speciation are called orthologues (Hillis and Moritz, 1990). Orthologues frequently have identical functions. To study groups of similar proteins within a genome, they first need to be clustered into families of paralogues. The choice of clustering method depends on the set of proteins and what the purpose of the clustering is. If it is mainly to get an idea of the

[1] To whom correspondence should be addressed. Telephone: +1-301-496-2477 Ext. 305. Fax: +1-301-480-9241. E-mail: esr@ncbi.nlm.nih.gov, rd@sanger.ac.uk.

[2] Present address: National Center for Biotechnology Information, National Library of Medicine, Building 38A, Room 8N805, National Institutes of Health, Bethesda, Maryland 20894.

0888-7543/97 $25.00

number of clusters, particularly in prokaryotes, a simple clustering method might give the best approximation. If one wants to build useful multiple alignments for each cluster, the clustering algorithm has to be able to infer domain boundaries and must split the sequences at these locations. Since we wanted to analyze the clusters by multiple alignment methods and since *C. elegans* has many multidomain proteins, we chose to use the Domainer algorithm (Sonnhammer and Kahn, 1994), which strives to perform a domainwise clustering. We selected the largest of the paralogue clusters generated this way that appeared to be unique to nematodes for more detailed analysis. In some cases, this resulted in a tentative functional assignment.

Finally, we compared the *C. elegans* proteins to the proteins in the completely sequenced genomes of *Saccharomyces cerevisiae* and *Haemophilus influenzae* and to a set of complete human proteins, to investigate the amount of conservation throughout the three kingdoms Eubacteria, Fungi, and animals, and to examine how useful knowledge of the *C. elegans* genome will be for understanding human biology.

MATERIALS AND METHODS

The Wormpep database contains all proteins predicted in the sequence produced by the *C. elegans* genome sequencing project and is available by anonymous FTP at ftp.sanger.ac.uk in /pub/databases/wormpep. The data are also in principle available in the EMBL and GenBank databases as cosmid DNA sequences and in SwissProt and PIR as proteins. There are, however, a number of reasons to base this analysis on Wormpep. The protein predictions are more up to date, since they are extracted directly from the latest version of ACEDB (Durbin and Thierry-Mieg, 1996). During this process a number of quality control checks are carried out to remove erroneous predictions. For example, genes that span two cosmids are correctly represented in ACEDB, but are not complete in the EMBL/GenBank sequence entries. A few proteins in Wormpep may still be fragments if they span two cosmids of which only one has been sequenced, or where the gene prediction is incorrect, or at the dozen or so boundaries between regions sequenced in Cambridge and St. Louis. The *C. elegans* World Wide Web servers (http://www.sanger.ac.uk and http://genome.wustl.edu/gsc) are the most up to date sources of sequence data, but only at the DNA level. Of the 7299 proteins in Wormpep 11, 36 are alternatively spliced variants of other genes confirmed by cDNA expressed sequence tag (EST) sequences. Often the difference between alternatively spliced genes is just one exon, coding for a few tens of amino acids. For this reason, only the first listed splicing variant of each gene was used for the analyses.

Blastp (Altschul *et al.*, 1990) was used with the BLOSUM62 substitution matrix. Blastp output was filtered by MSPcrunch to remove biased composition matches and enhance the selection of consistent multiple match segments (Sonnhammer and Durbin, 1994).

The Pfam matching was performed with the hmmfs and hmmls search programs, which are part of the HMMER package (Eddy, 1995).

The clustering of Wormpep was performed by Version 1.6 of the Domainer program (Sonnhammer and Kahn, 1994), using pairwise homology information from Blastp. A score threshold of 90 was used, and MSPcrunch was run in a mode that trims off overlapping ends of consecutive matches. Pfam-A 1.0 was used for the clustering analysis. The clusters were analyzed for similarity outside Nematoda by searching the consensus sequence of each cluster against the NCBI nonredundant database as of August 8, 1997, using Tblastn and MSPcrunch and removing matches to Nematoda with the program tax_filt (Walker and Koonin, 1997).

The largest apparently nematode-specific families were analyzed by running HMMs derived from the multiple alignments against swir11, which is a non-redundant combination of Wormpep 11, SwissProt 33, and SwissProt-TREMBL 47β (Bairoch and Apweiler, 1997). Prosite (Bairoch *et al.*, 1997) patterns were searched with the perl script queryprosite, and coiled coil predictions were made with the program Pepcoil, which is part of the EGCG package (Rice *et al.*, 1995) and uses the algorithm by Lupas *et al.* (1991). The multiple alignments and the tree were generated with Clustalw (Thompson *et al.*, 1994). The alignment figures were produced with Belvu (E. Sonnhammer, unpublished), and the tree figure with TreeTool (Maidak *et al.*, 1997).

The protein sets for the pairwise genome to genome comparisons were assembled the following way. *Homo sapiens,* all entries in SwissProt 33; *C. elegans,* all entries in Wormpep 11, except alternatively spliced versions of the same gene; *S. cerevisiae,* all entries in SwissProt 33 and all entries in SwissProt-TREMBL that were not 100% identical to (a part of) a SwissProt entry. To ascertain that no *S. cerevisiae* proteins were missed in Table 5 of proteins unique to *H. influenzae* and *C. elegans,* all candidate proteins were also compared to the *S. cerevisiae* DNA sequences in EMBL 48 using Tfasta and Tblastn; *H. influenzae,* all entries in the TIGR set (ftp://ftp.tigr.org/pub/data/h_influenzae). The human and yeast datasets contained both nuclear and mitochondrial encoded proteins. The 13 mitochondrial *C. elegans* proteins in SwissProt 33 were not included. The *S. cerevisiae* dataset was somewhat redundant even after excluding the 100% matching and included sequences. We did not wish to remove less than 100% identical proteins on the basis of similarity only, to avoid removing very similar paralogues. The human dataset could have been augmented by using EST data. We chose to not use these data, since their fragmentary nature makes the estimate of the number of matches and their extent uncertain.

For the pairwise genome to genome comparisons, the MSPcrunch parameters were set more stringently than the default, to reduce the number of spurious matches. We raised the score range of the "twilight zone" from 35–75 to 45–80 and the bias composition criterion to 0.8 with no pseudocounts. The accuracy was assessed by manual inspection of a few genome comparisons in Blixem (Sonnhammer and Durbin, 1994) and Dotter (Sonnhammer and Durbin, 1996). In the *C. elegans* to *H. sapiens* comparison, 125 protein assignments were removed by the increase in MSPcrunch stringency. Of these, only 9 were found likely to be true matches. We also performed the same analysis on the 150 assignments (2% of the *C. elegans* proteins) in the *C. elegans* to *S. cerevisiae* comparison that had only matches scoring below 80. Of these, only about 10 were dubious. Our method should thus be a good compromise between sensitivity and selectivity for genomic comparison purposes. An alternative approach would be to apply other types of programs for postprocessing matches in the twilight zone, such as dynamic programming and multiple alignment methods (Koonin *et al.*, 1996b; Tatusov *et al.*, 1996). MSPcrunch could also have been used in a less stringent mode, combined with manual processing of twilight zone matches. For a detailed analysis,

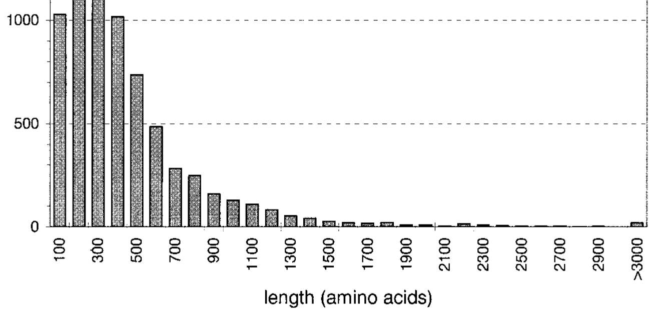


FIG. 1. Length distribution of Wormpep 11 entries.

TABLE 1

The Occurrence of the Most Frequent Pfam Domains ($n \geq 10$) in Wormpep 11, Which Comprises about Half of the Proteins in *Caenorhabditis elegans*, and the Number of Members of the Same Domain Families in the Entire Yeast Genome

WP11 domains/proteins	Yeast domains/proteins	Pfam Accession No.	Pfam annotation
219/203	158/136	PF00069	Eukaryotic protein kinase domain
185/40	46/20	PF00023	Ank repeat
159/66	96/45	PF00096	Zinc finger, C2H2 type
135/32	1/1	PF00008	EGF-like domain
125/21	2/2	PF00041	Fibronectin type III domain
89/21	0/0	PF00047	IG superfamily
81/72	0/0	PF00001	7 transmembrane receptor (rhodopsin family)
81/21	0/0	PF00090	Thrombospondin type 1 domain
80/32	112/52	PF00400	WD domain, Gβ repeats
76/49	88/50	PF00076	RNA recognition motif (aka RRM, RBD, or RNP domain)
69/14	0/0	PF00057	Low-density lipoprotein receptor domain class A
60/58	0/0	PF00105	Zinc finger, C4 type (two domains)
60/41	0/0	PF00065	Neurotransmitter-gated ion channel
59/16	0/0	PF00014	Kunitz/bovine pancreatic trypsin inhibitor domain
53/23	19/9	PF00036	EF hand
47/46	8/8	PF00046	Homeobox domain
46/5	0/0	PF00028	Cadherin
46/29	54/33	PF00005	ABC transporters
42/31	7/3	PF00102	Protein-tyrosine phsophatase
40/3	0/0	PF00435	Spectrin α chain, repeated domain
38/5	0/0	PF00053	Laminin EGF-like (domains III and V)
34/22	14/14	PF00149	Ser/Thr protein phosphatases
33/26	0/0	PF00201	UDP-glucoronosyl and UDP-glucosyl transferases
33/26	0/0	PF00059	Lectin C-type domain short and long forms
32/31	7/7	PF00099	Zinc-binding metalloprotease domain
32/15	0/0	PF00431	CUB domain
30/9	14/4	PF00013	KH domain family of RNA binding proteins
30/30	13/12	PF00106	Alcohol/other dehydrogenases, short chain type
30/23	27/23	PF00018	Src homology domain 3
28/28	0/0	PF00104	Ligand-binding domain of nuclear hormone receptors
27/27	11/11	PF00125	Core histones H2A, H2B, H3, and H4
32/32	82/80	PF00271	Helicases conserved C-terminal domain
26/26	17/17	PF00097	Zinc finger, C3HC4 type (RING finger)
25/24	5/5	PF00010	Helix–loop–helix DNA binding domain
25/17	4/3	PF00412	LIM domain-containing proteins
24/24	4/4	PF00067	Cytochrome P450
24/21	37/36	PF00153	Mitochondrial carrier proteins
24/16	10/5	PF00168	C2 domain
23/10	4/1	PF00520	Ion transport proteins
22/19	1/1	PF00017	Src homology domain 2
20/20	27/27	PF00071	Ras family (contains ATP/GTP binding P-loop)
20/17	0/0	PF00211	Guanylate cyclases
20/12	0/0	PF00135	Carboxylesterases
18/18	43/43	PF00083	Sugar (and other) transporters
18/18	8/8	PF00043	Glutathione *S*-transferases
18/17	8/8	PF00078	Reverse transcriptase (RNA-dependent DNA polymerase)
17/4	0/0	PF00084	Sushi domain
17/17	20/20	PF00169	PH (pleckstrin homology) domain
17/16	3/3	PF00188	SCP-like extracellular proteins
16/9	16/11	PF00439	Bromodomain
16/7	0/0	PF00054	Laminin G domain
16/13	0/0	PF00092	von Willebrand factor type A domain
16/11	12/10	PF00085	Thioredoxins
15/13	2/1	PF00130	Phorbol esters/diacylglycerol binding domain
15/13	33/28	PF00004	ATPases associated with various cellular activities (AAA)
14/4	10/6	PF00240	Ubiquitin family
14/3	0/0	PF00058	Low-density lipoprotein receptor domain class B
13/9	8/6	PF00225	Kinesin motor domain
13/9	0/0	PF00038	Intermediate filament proteins
13/13	19/18	PF00226	DnaJ, prokaryotic heat shock protein
12/6	6/4	PF00450	Serine carboxypeptidases
12/12	10/10	PF00501	AMP-binding enzymes

TABLE 1—Continued

WP11 domains/proteins	Yeast domains/proteins	Pfam Accession No.	Pfam annotation
12/12	27/27	PF00270	DEAD and DEAH box helicases
12/11	7/7	PF00328	Histidine acid phosphatases
11/8	0/0	PF00335	4 transmembrane segments integral membrane proteins
11/8	0/0	PF00060	Ligand-gated ionic channels
11/7	0/0	PF00337	Vertebrate galactoside-binding lectins
11/7	2/2	PF00282	Pyridoxal-dependent decarboxylases conserved domain
11/11	2/2	PF00503	G-protein α subunit
10/9	6/6	PF00433	Protein kinase C terminal domain
10/9	5/5	PF00248	Aldo/keto reductase family
10/9	5/5	PF00063	Myosin head (motor domain)
10/10	15/15	PF00442	Ubiquitin carboxyl-terminal hydrolases family 2
10/10	12/12	PF00179	Ubiquitin-conjugating enzymes

Note. The number of domains may be somewhat overestimated for some families due to multiple fragment matches, and because multiple alternative splicing products were included, the number of *C. elegans* proteins may be slightly too high.

manual inspection of the results is essential. However, the accuracy achieved by MSPcrunch without manual processing in the twilight zone seems adequate for a reasonably reliable estimate of the overall percentage similarity between genomes. If anything, our method is conservative; we accept missing some weak matches as a tradeoff for rejecting most spurious ones.

The smaller sets of *C. elegans* proteins for Figs. 3 and 10 were generated by selecting a random subset from Wormpep 11. The regression was performed by fitting a logarithmic function to the simulated datapoints in Microsoft Excel.

RESULTS

The *C. elegans* proteins for this analysis were predicted from genomic sequences, as compiled in Wormpep, Release 11 (see Materials and Methods). The distribution of protein lengths in Wormpep is very skewed, as seen in Fig. 1. The mean length is 450, while the median is only 342. This is due to a small number of very long proteins. Nineteen predicted proteins have more than 3000 amino acid residues. The largest protein so far is K07E12.1, with 13,055 residues. It contains some 10 fibronectin type 3 domains, 6 immunoglobulin superfamily domains (cell-adhesion molecule-like), 1 epidermal growth factor-like domain, 3 von Willebrand factor type A domains, and about 60 repeats of a new type. Such multiple-domain giants are frequently extracellular proteins that often have a role in cell–cell binding.

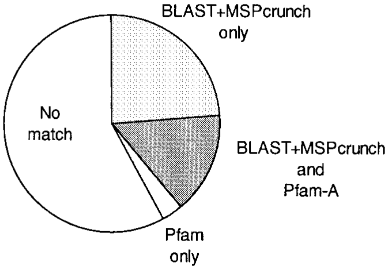


FIG. 2. Fractions of *C. elegans* proteins that can be annotated based on homology, using the pairwise BLAST+MSPcrunch approach (Sonnhammer and Durbin, 1994) and the family-based Pfam approach. Together they add up to an annotation level of about 43%.

The accuracy of the gene predictions in Wormpep depends on the amount of evidence available. The predictions were made in an integrated analysis/gene prediction workbench (Sonnhammer and Durbin, 1994), in which an annotator combines different types of evidence (Scharf *et al.*, 1994; Zhang *et al.*, 1994; Bisson and Garreau, 1995; Rubin, 1996). Genes for which ESTs have been sequenced can be considered experimentally verified in the regions that match, and genes with strong similarity to other proteins are usually close to 100% correct. Gene predictions that lack these extrinsic pieces of evidence must rely solely on the DNA sequence, exploiting statistical evidence for coding potential and splicing signals. This is the case for a minority of genes, however, since about one-third of the genes have at least one EST match, and over half are similar to other proteins from *C. elegans* or other organisms. The program that has been used for gene prediction, Genefinder (P. Green, unpublished), generally predicts most of the exons in the middle of genes correctly, while exons at the start and end, which often contain weaker signals, frequently are mispredicted if no extrinsic evidence is available. Occasionally, close neighboring genes may be fused, and single genes with long introns may be fragmented. Because of the uncertainty in the gene predictions, there is a certain error margin in the results below.

Classification of Wormpep Entries by Pfam

Of the 7299 proteins in Wormpep 11, 2868 (39%) are functionally annotated. This was done manually, using a computer-assisted analysis workbench built around ACEDB (Sonnhammer and Durbin, 1994). This annotation is not always easy to use for summary purposes, because the nomenclature used is variable, and it is not always complete. For example, over 20% of the eukaryotic protein kinases found by Pfam did not have the word "kinase" in the annotation. About half of these lacked annotation completely, while the other half had other annotations, such as "receptor" or "cell division

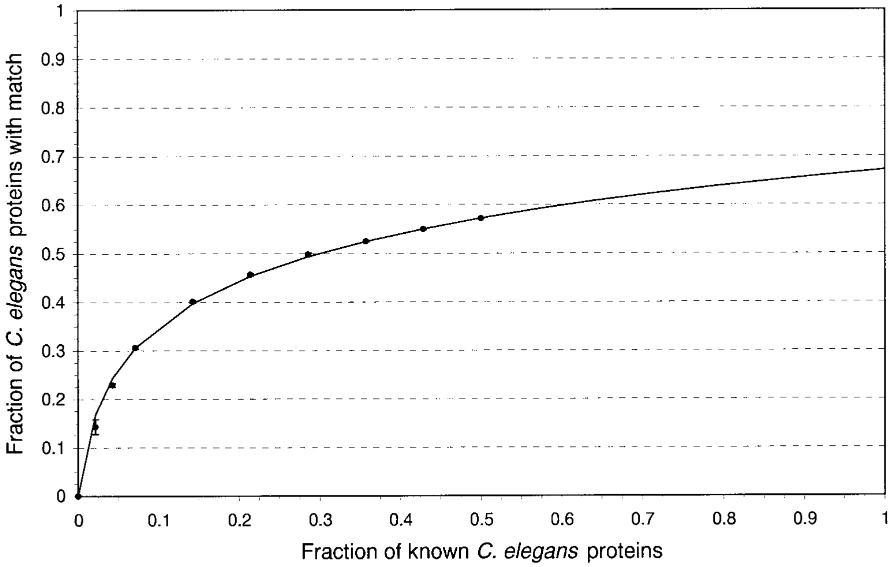


FIG. 3. Projection of the fraction of proteins that match one another within the *C. elegans* genome, for different fractions of known *C. elegans* proteins. The datapoints below 50% were simulated by taking fractions of the currently known *C. elegans* proteins in Wormpep 11. The values are averages from three independent experiments, and the error bars are standard deviations.

control protein." Guanylate cyclases also match the protein kinase family.

To summarize the families in a more consistent fashion, we used the Pfam database of protein domain families (Sonnhammer *et al.*, 1997). We compared all Wormpep 11 sequences to all Pfam 2.0 families, using as significance cutoffs Pfam's previously recorded family-specific cutoffs that were chosen to exclude negatives. All protein domains with 10 or more examples are listed in Table 1. Many of the most frequent domains are multiply repeated in single proteins. For example, 38 laminin-type EGF domains are spread in only 5 pro-

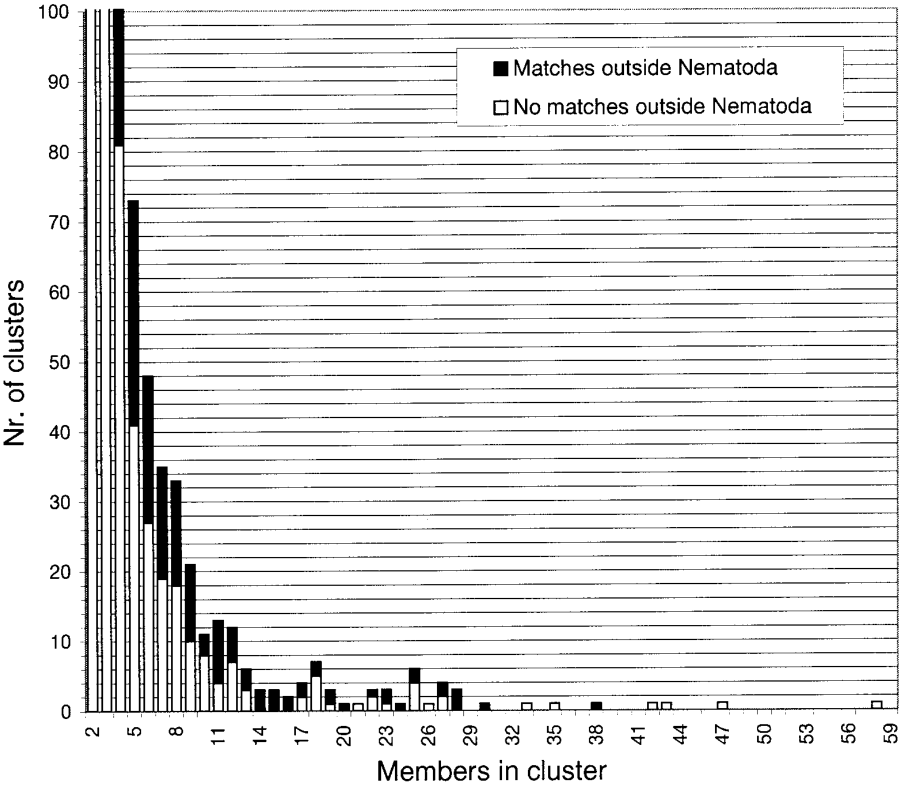


FIG. 4. Histogram of Wormpep 11 cluster sizes, produced by Domainer clustering of all segments not matching known protein families in Pfam. The consensus sequence of each cluster was searched against the NR database to determine whether the cluster is unique to Nematoda.

TABLE 2

Apparently Nematode-Specific Protein Domain Families

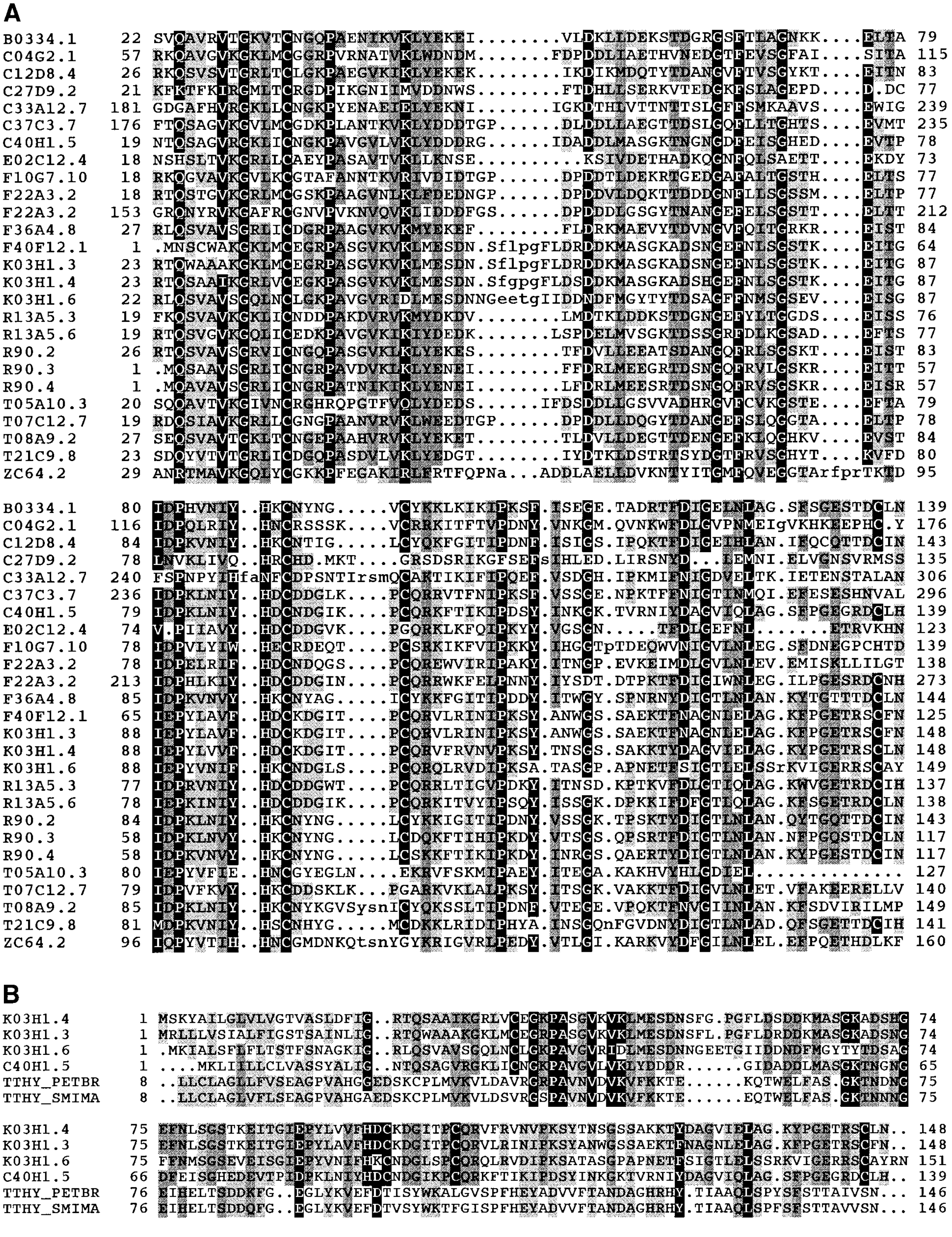
Family	Members (No. of domains in parentheses)	Alternative name[a]	Domains	Proteins	Length	Putative function
1	C18H2.1 C18H2.3(2) C18H2.4 F37A4.4 F56D5.9		6	5	1000	? (Contains one ank repeat)
2	B0334.1 C04G2.1 C12D8.4 C14C10.2 C27D9.2 C33A12.7 C37C3.7 C40H1.5 E02C12.4(2) F10G7.10 F22A3.2(2) F26G1.3 F36A4.8 F40F12.1 K03H1.3 K03H1.4 K03H1.6 R13A5.3 R13A5.6 R90.2 R90.3 R90.4 T05A10.3 T07C12.7 T07C4.5 T08A9.2(2) T14G10.3 T14G10.4 T21C9.89 ZC64.2		33	30	160	Hormone transporter
3	B0244.4 B0244.5 B0244.6(3) B0244.7(3) ZK418.6 ZK418.7		9	9	195	? (transmembrane)
4	C14A4.10 C18F10.4 C18F10.5 C18F10.6 C18F10.8 C33A12.10 C33A12.11 C33A12.8 C33D9.4 C34C6.1 F48D6.2 R07B5.6 R13F6.3 T01B7.2 T04A8.1 T04A8.2 T12A2.10 T12A2.11 T12A2.12 T12A2.13 T12A2.9 T13A10.13 T19C4.3 T21C9.7 T23F11.5	srg	25	25	400	GPCR
5	AH6.10 Ah6.11 AH6.12 AH6.13(2) AH6.14 AH6.4 AH6.6 AH6.7 AH6.8 AH6.9 B0304.5(2) B0304.6 B0304.7(3) C27D6.10 C27D6.6 C27D6.7 C27D6.8 C27D6.9(2) C33G8.5 C56C10.5 F18C5.1(2) F18C5.6 F18C5.8 F23F12.10 F37C12.15 F37C12.16 F44F4.13 F44F4.5 F44F4.7 F49E12.5 F58A6.10 F5A6.11 F58A6.6 K11E4.4 RO4B5.10 RO5H5.6 R10H1.2 T11A5.3 T11A5.4(2) T19D12.8 T21H8.2 T21H8.3	sra	43	42	335	GPCR
6	B0228.3 (13)		13	1	230	?
7	F26C11.3(9)		9	1	80	?
8	B0564.3 B0564.4 C07A9.8 C09B9.3(3) C29F4.2 F32G8.4 R13.3 T19C3.1 T20G5.4 ZC518.1 ZK675.3 ZK688.2		13	12	386	? (transmembrane)
9	B0547.3 C06C3.6 C06G8.4 C12D8.12 C33G8.1 C39H7.6(2) C42D4.12C42D4.4 C42D4.5 C42D4.9 C45B11.4 C48C5.1 C50C10.6 C53B7.5 C54A12.2 D1054.12 F13G3.2 F15A2.4 F17A2.10 F17A2.11 F17A2.12 F17A2.6 F17A2.7 F17A2.8 F17A2.9 F18E3.5 F28H7.1 F32G8.1 F33H1.5 F40F9.4 F47G9.2 F52D2.7 F57A8.3 F58G4.5 F58G4.6 F58G4.7 K02A2.1 K02A2.2 M7.9 R04B5.8 R04D3.6 R04D3.7 R04D3.8 R05H5.1 R07B5.1 R07B5.2 R08C7.7 R09F10.6 R11D1.5 R11D1.6 T07C12.1 T07C12.4 T07C12.5 T08H10.2 T18H9.4 T19E7.5 T22H6.3 T22H6.4 ZK829.8	srd	60	59	315	GPCR
10	K07E12.1(58)		58	1	195	?

Note. GPCR, G-protein-coupled receptor.

[a] These families have been found independently and are described elsewhere (Troemel *et al.*, 1995).

teins, and the 184 ankyrin repeats are in only 40 proteins. The collagen family, which is one of the most abundant *C. elegans* protein families, is not listed in Table 1 because it was not included in Pfam 2.0. Also listed in Table 1 is the occurrence of these domains in the entire *S. cerevisiae* genome. Many proteins specific for metazoa are completely absent in yeast, while many occur at approximately the same rate. The only family that appears to be clearly more frequent in yeast is the helicases (Pfam PF00270 and PF00271).

To look specifically for novel Pfam classifications, the 4431 proteins in Wormpep 11 that had no functional annotation after analysis with the standard Blast/MSPcrunch approach (Sonnhammer and Durbin, 1994) were searched for Pfam matches. Five hundred nineteen matches to 319 previously unannotated *C. elegans* sequences were found. A number of these matches had very high scores, indicating that they should have been found by BLAST too. We have found empirically that most matches found by Pfam but not by BLAST have scores below approximately 35 bits. Roughly half of the matches scored lower than this, thus representing genuinely novel classifications that are likely to have been missed because the similarity to any one other protein was too weak. Matches above this score are likely to have been missed due to human error.



A

```
B0334.1     22 SVQAVRVTGKVTCNGQPAENIKVKLYEKEI........VLDKLLDEKSTDGRGSFTLAGNKK....ELTA 79
C04G2.1     57 RKQAVGVKGKLMCGGRPVRNATVKLWDNDM.......FDPDDLIAETHVNEDGTFEVSGFAI....SITA 115
C12D8.4     26 RKQSVSVTGRLTCLGKPAEGVKIKLYEKEK........IKDIKMDQTYTDANGVFTVSGYKT....EITN 83
C27D9.2     21 KFKTFKIRGMLTCRGDPIKGNIIMVDDNWS........FTDHLLSERKVTEDGKFSLAGEPD....D.DC 77
C33A12.7   181 GDGAFHVRGKILCNGKPYENAEIELYEKNI.......IGKDTHLVTTNTTSLGFFSMKAAVS....EWIG 239
C37C3.7    176 FTQSAGVKGVIMCGDKPLANTKVKLYDDDTGP......DLDDLLAEGTTDSLGQFLLTGHTS....EVMT 235
C40H1.5     19 NTQSAGVRGKLICNGKPAVGVLVKLYDDDRG......IDADDLMASGKTNGNGDFEISGHED....EVTP 78
E02C12.4    18 NSHSLTVKGRLLCAEYPASAVTVKLLKNSE..........KSIVDETHADKQGNFQLSAETT....EKDY 73
F10G7.10    18 RKQGVAVKGVLKCGTAFANNTKVRIVDIDTGP......DPDDTLDEKRTGEDGAFALTGSTH....ELTS 77
F22A3.2     18 RTQSTGVKGRLMCGSKPAAGVNLKLFDEDNGP......DPDDVLDQKTTDDDGNFLLSGSSM....ELTP 77
F22A3.2    153 GRQNYRVKGAFRCGNVPVKNVQVKLIDDDFGS......DPDDDLGSGYTNANGEFELSGSTT....ELTT 212
F36A4.8     27 RLQSVAVSGRLICDGRPAAGVKVKMYEKEF........FLDRKMAEVYTDVNGVFQITGRKR....EIST 84
F40F12.1     1 .MNSCWAKGKLMCEGRPASGVKVKLMESDN.SflpgFLDRDDKMASGKADSNGEFNLSGSTK....EITG 64
K03H1.3     23 RTQWAAAKGKLMCEGRPASGVKVKLMESDN.SflpgFLDRDDKMASGKADSNGEFNLSGSTK....EITG 87
K03H1.4     23 RTQSAAIKGRLVCEGKPASGVKVKLMESDN.SfgpgFLDSDDKMASGKADSHGEFNLSGSTK....EITG 87
K03H1.6     22 RLQSVAVSGQLNCLGKPAVGVRIDLMESDNNGeetgIIDDNDFMGYTYTDSAGFFNMSGSEV....EISG 87
R13A5.3     19 FKQSVAVKGKLICNDDPAKDVRVKMYDKDV........LMDTKLDDKSTDGNGEFYLTGGDS....EISS 76
R13A5.6     19 RTQSVGVKGQLICEDKPAVGVKIKIYDEDK.......LSPDELMVSGKTDSSGRFDLKGSAD....EFTS 77
R90.2       26 RTQSVAVSGRVICNGQPASGVKIKLYEKES........TFDVLLEEATSDANGQFRLSGSKT....EIST 83
R90.3        1 .MQSAAVSGRLICNGRPAVDVKLKLYENEI........FFDRLMEEGRTDSNGQFRVLGSKR....EITT 57
R90.4        1 .MQAVAVSGRLICNGRPATNIKIKLYENEI........LFDRLMEESRTDSNGQFRVSGSKR....EISR 57
T05A10.3    20 SQQAVTVKGIVNCRGHRQPGTFVQLYDEDS......IFDSDDLLGSVVADHRGVFCVKGSTE....EFTA 79
T07C12.7    19 RDQSIAVKGRLLCGNGPAANVRVKLWEEDTGP......DPDDLLDQGYTDANGEFSLQGGTA....ELTP 78
T08A9.2     27 SEQSVAVTGKLTCNGEPAAHVRVKLYEKET........TLDVLLDEGTTDENGEFKLQGHKV....EVST 84
T21C9.8     23 SDQYVTVTGRLICDGQPASDVLVKLYEDGT........IYDTKLDSTRTSYDGTFRVSGHYT....KVFD 80
ZC64.2      29 ANRTMAVKGQLYCGKKPFEGAKIRLFRTFQPNa...ADDLAELLDVKNTYITGMFQVEGGTArfprTKTD 95

B0334.1     80 IDPHVNIY..HKCNYNG.....VCYKKLKIKIPKSF.ISEGE.TADRTFDIGELNLAG.SFSGESTDCLN 139
C04G2.1    116 IDPQLRIY..HNCRSSSK....VCRRKITFTVPDNY.VNKGM.QVNKWFDLGVPNMEIgVKHKEEPHC.Y 176
C12D8.4     84 IDPKVNIY..HKCNTIG.....LCYQKFGITIPDNF.ISIGS.IPQKTFDIGEIHLAN.IFQCQTTDCIN 143
C27D9.2     78 LNVKLIVQ..HRCHD.MKT....GRSDSRIKGFSEFsIHLED.LIRSNYD...IEMNI.ELVGNSVRMSS 135
C33A12.7   240 FSPNPYIHfaNFCDPSNTIrsmQCAKTIKIFIPQEF.VSDGH.IPKMIFNIGDVELTK.IETENSTALAN 306
C37C3.7    236 IDPKLNIY..HDCDDGLK....PCQRRVTFNIPKSF.VSSGE.NPKTFFNIGTINMQI.EFESESHNVAL 296
C40H1.5     79 IDPKLNIY..HDCNDGIK....PCQRKFTIKIPDSY.INKGK.TVRNIYDAGVIQLAG.SFPGEGRDCLH 139
E02C12.4    74 V.PIIAVY..HDCDDGVK....PGQRKLKFQIPKYY.VGSGN.....TFDLGEFNL.......ETRVKHN 123
F10G7.10    78 IDPVLYIW..HECRDEQT....PCSRKIKFVIPKKY.IHGGTpTDEQWVNIGVLNLEG.SFDNEGPCHTD 139
F22A3.2     78 IDPELRIF..HDCNDQGS....PCQREWVIRIPAKY.ITNGP.EVKEIMDLGVLNLEV.EMISKLLILGT 138
F22A3.2    213 IDPHLKIY..HDCDDGIN....PCQRRWKFEIPNNY.IYSDT.DTPKTFDIGIWNLEG.ILPGESRDCNH 273
F36A4.8     85 IDPKVNVY..HKCNYAG.....ICYKKFGITIPDDY.ITWGY.SPNRNYDIGTLNLAN.KYTGTTTDCLN 144
F40F12.1    65 IEPYLAVF..HDCKDGIT....PCQRVLRINIPKSY.ANWGS.SAEKTFNAGNIELAG.KFPGETRSCFN 125
K03H1.3     88 IEPYLAVF..HDCKDGIT....PCQRVLRINIPKSY.ANWGS.SAEKTFNAGNLELAG.KFPGETRSCFN 148
K03H1.4     88 IEPYLVVF..HDCKDGIT....PCQRVFRVNVPKSY.TNSGS.SAKKTYDAGVIELAG.KYPGETRSCLN 148
K03H1.6     88 IEPYVNIF..HKCNDGLS....PCQRQLRVDIPKSA.TASGP.APNETFSIGTLELSSrKVIGERRSCAY 149
R13A5.3     77 IDPRVNIY..HDCDDGWT....PCQRRLTIGVPDKY.ITNSD.KPTKVFDLGTIQLAG.KWVGETRDCIH 137
R13A5.6     78 IEPKINIY..HDCDDGIK....PCQRKITVYIPSQY.ISSGK.DPKKIFDFGTIQLAG.KFSGETRDCLN 138
R90.2       84 IDPKLNIY..HKCNYNG.....LCYKKIGITIPDNY.VSSGK.TPSKTYDIGTLNLAN.QYTGQTTDCIN 143
R90.3       58 IDPKLNVY..HKCNYNG.....LCDQKFTIHIPKDY.VTSGS.QPSRTFDIGTLNLAN.NFPGQSTDCLN 117
R90.4       58 IDPKVNVY..HKCNYNG.....LCSKKFTIKIPKDY.INRGS.QAERTYDIGTLNLAN.KYPGESTDCIN 117
T05A10.3    80 IEPYVFIE..HNCGYEGLN....EKRVFSKMIPAEY.ITEGA.KAKHVYHLGDIEL.............. 127
T07C12.7    79 IDPVFKVY..HKCDDSKLK...PGARKVKLALPKSY.ITSGK.VAKKTFDIGVLNLET.VFAKEERELLV 140
T08A9.2     85 IDPKLNIY..HKCNYKGVSysnICYQKSSLTIPDNF.VTEGE.VPQKTFNVGIINLAN.KFSDVIRILMP 149
T21C9.8     81 MDPKVNIY..HSCNHYG.....MCDKKLRIDIPHYA.INSGQnFGVDNYDIGTLNLAD.QFSGETTDCIH 141
ZC64.2      96 IQPYVTIH..HNCGMDNKQtsnYGYKRIGVRLPEDY.VTLGI.KARKVYDFGILNLEL.EFPQETHDLKF 160
```

B

```
K03H1.4      1 MSKYAILGLVLVGTVASLDFIG..RTQSAAIKGRLVCEGKPASGVKVKLMESDNSFG.PGFLDSDDKMASGKADSHG 74
K03H1.3      1 MRLLLVSIALFIGSTSAINLIG..RTQWAAAKGKLMCEGRPASGVKVKLMESDNSFL.PGFLDRDDKMASGKADSNG 74
K03H1.6      1 .MKIALSFLFLTSTFSNAGKIG..RLQSVAVSGQLNCLGKPAVGVRIDLMESDNNGEETGIIDDNDFMGYTYTDSAG 74
C40H1.5      1 ....MKLIILLCLVASSYALIG..NTQSAGVRGKLICNGKPAVGVLVKLYDDDR......GIDADDLMASGKTNGNG 65
TTHY_PETBR   8 ..LLCLAGLLFVSEAGPVAHGGEDSKCPLMVKVLDAVRGRPAVNVDVKVFKKTE......KQTWELFAS.GKTNDNG 75
TTHY_SMIMA   8 ..LLCLAGLVFLSEAGPVAHGAEDSKCPLMVKVLDSVRGSPAVNVDVKVFKKTE......EQTWELFAS.GKTNNNG 75

K03H1.4     75 EFNLSGSTKEITGIEPYLVVFHDCKDGITPCQRVFRVNVPKSYTNSGSSAKKTYDAGVIELAG.KYPGETRSCLN.. 148
K03H1.3     75 EFNLSGSTKEITGIEPYLAVFHDCKDGITPCQRVLRINIPKSYANWGSSAEKTFNAGNLELAG.KFPGETRSCFN.. 148
K03H1.6     75 FFNMSGSEVEISGIEPYVNIFHKCNDGLSPCQRQLRVDIPKSATASGPAPNETFSIGTLELSSRKVIGERRSCAYRN 151
C40H1.5     66 DFEISGHEDEVTPIDPKLNIYHDCNDGIKPCQRKFTIKIPDSYINKGKTVRNIYDAGVIQLAG.SFPGEGRDCLH.. 139
TTHY_PETBR  76 EIHELTSDDKFG..EGLYKVEFDTISYWKALGVSPFHEYADVVFTANDAGHRHY.TIAAQLSPYSFSTTAIVSN... 146
TTHY_SMIMA  76 EIHELTSDDQFG..EGLYKVEFDTVSYWKTFGISPFHEYADVVFTANDAGHRHY.TIAAQLSPFSFSTTAVVSN... 146
```

FIG. 5. (**A**) Alignment of a selection of the members in apparently nematode-specific family 2. (**B**) Alignment of selected members to transthyretins (SwissProt P49142 and P49143). Although tentative, the similarity suggests that this may be a family of hormone transporters.

About 25% of all Wormpep proteins have at least one domain that matches a Pfam-A family. The matching regions are on average about half the length of the proteins, so about 13% of the residues in Wormpep are covered by Pfam-A. Since Pfam-A contains only the most common protein domain families, these numbers are necessarily much lower than the fraction of Wormpep that has matches found by all-protein searches using BLAST. Figure 2 illustrates the relative proportions of annotation in Wormpep 11. Overall, 39% of the proteins have functional annotation based on BLAST/MSPcrunch analysis. Some 4% of the Pfam-A matches were to previously unannotated proteins. Although Pfam-A currently

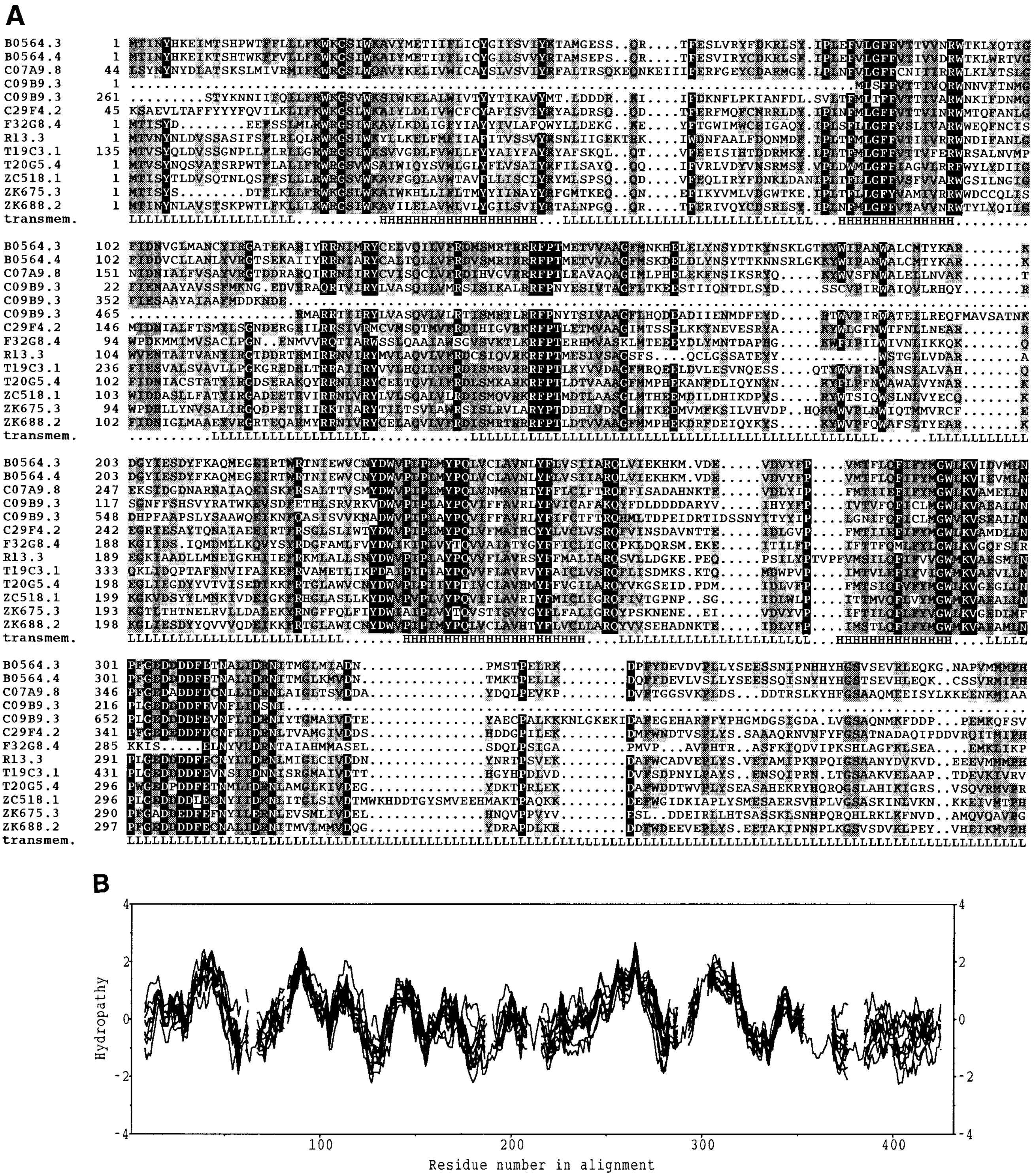


FIG. 6. (**A**) Alignment of all members of apparently nematode-specific family 8. The line 'transmem.' shows segments predicted to be membrane-spanning helices (H) and loops (L) by PHDhtm (Rost *et al.*, 1995). (**B**) Combined hydropathy plot of all members. No putative function has been assigned to this family, but the members are likely to have a membrane location.

adds only a few more percent to the fraction of annotated proteins, the analysis of proteins with BLAST matches benefits from Pfam matches too, by clearer indication of domains and family annotation. Furthermore, cases in which Pfam detected previously unidentified domains in previously annotated proteins are not reflected in Fig. 2. Considering that this analysis is based on only the second release of Pfam, already a substantial fraction of Wormpep is covered. The two approaches thus complement each other well; overall annotation improvement is achieved by combining them.

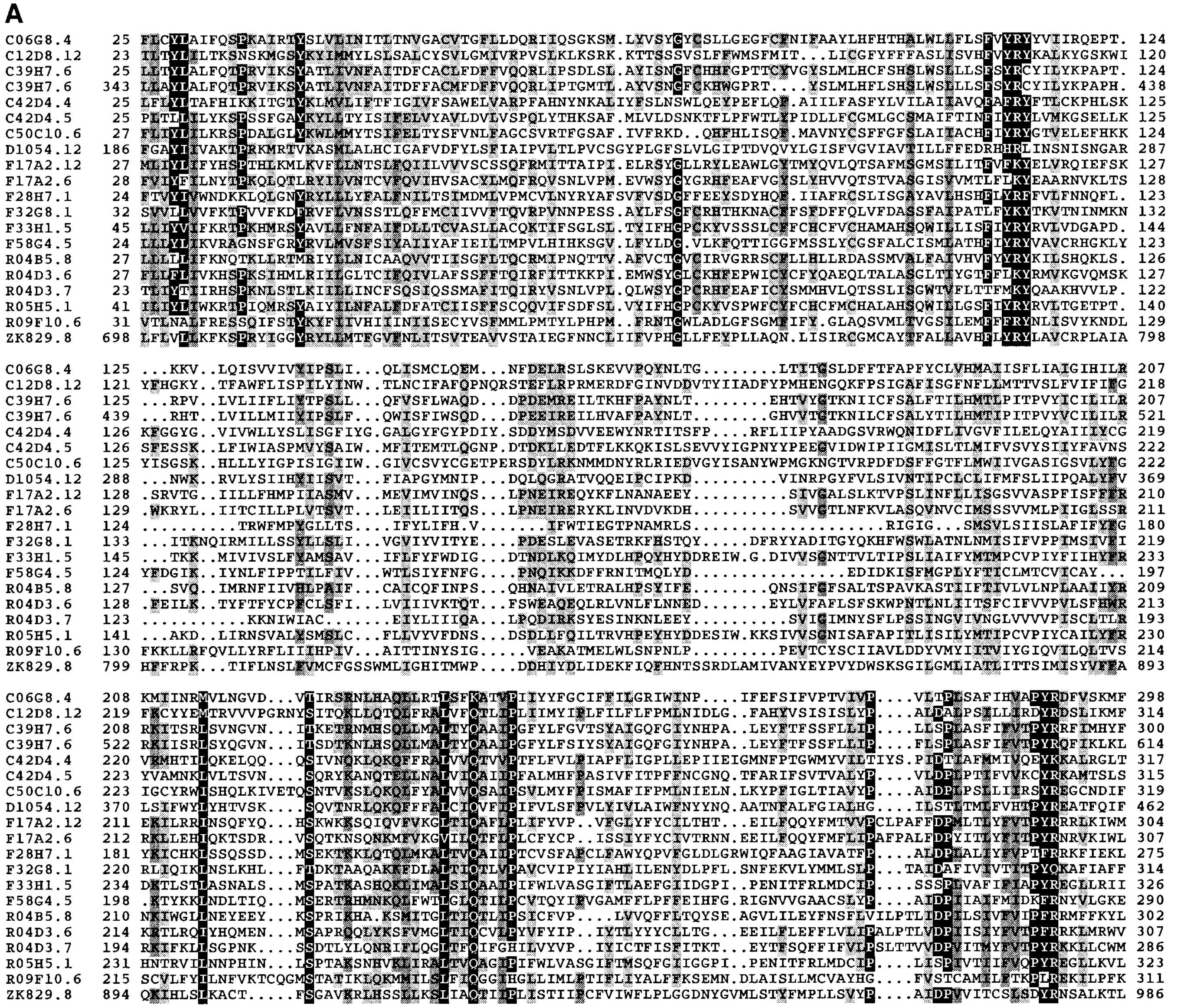


FIG. 7. (**A**) Alignment of a selection of the members in apparently nematode-specific family 9. This family has weak similarity to G-protein-coupled receptors, which is supported by the hydropathy profile (not shown). (**B**) The distribution of all members in apparently nematode-specific family 9 over the six chromosomes of *C. elegans* (black horizontal lines). The light-shaded areas represent the regions that have been cloned, and the dark-shaded areas are the sequenced regions that Wormpep 11 was based on. (**C**) Tree of a selection of members in this family, showing bootstrap values from 1000 trials. The most similar sequences are almost invariably close in the genome, either on the same cosmid (e.g., the proteins on F17A2) or on the neighbor cosmid (e.g., B0547 and C39H7, F33H1 and R05H5). The distance between cosmids F17A2 and R04D3 is approximately 0.3 Mb.

Clustering of Wormpep Proteins

Many Wormpep proteins match other Wormpep proteins, forming clusters of related proteins. The extent of internal matching in *C. elegans* is illustrated in Fig. 3, which shows the percentage of proteins that match another for the current fraction of the genome and for smaller fractions. This was measured in a relatively conservative way, counting only matches that are significant according to MSPcrunch run with stringent parameters (see Materials and methods). By extrapolation, we find that when all the sequences are available, about two-thirds of the *C. elegans* proteins should match another in the entire genome.

To examine the size distribution of paralogue families in Wormpep, a clustering analysis was performed. Many clustering methods are known (e.g., Romesburg, 1989), but only a few are well-suited for protein sequences. The two major problems are the fact that many sequences contain more than one protein domain and the uncertainty of family membership for marginal matches. At present, no clustering algorithm can solve all these problems without compromise. A choice has to be made between using a simple algorithm that is poorly suited

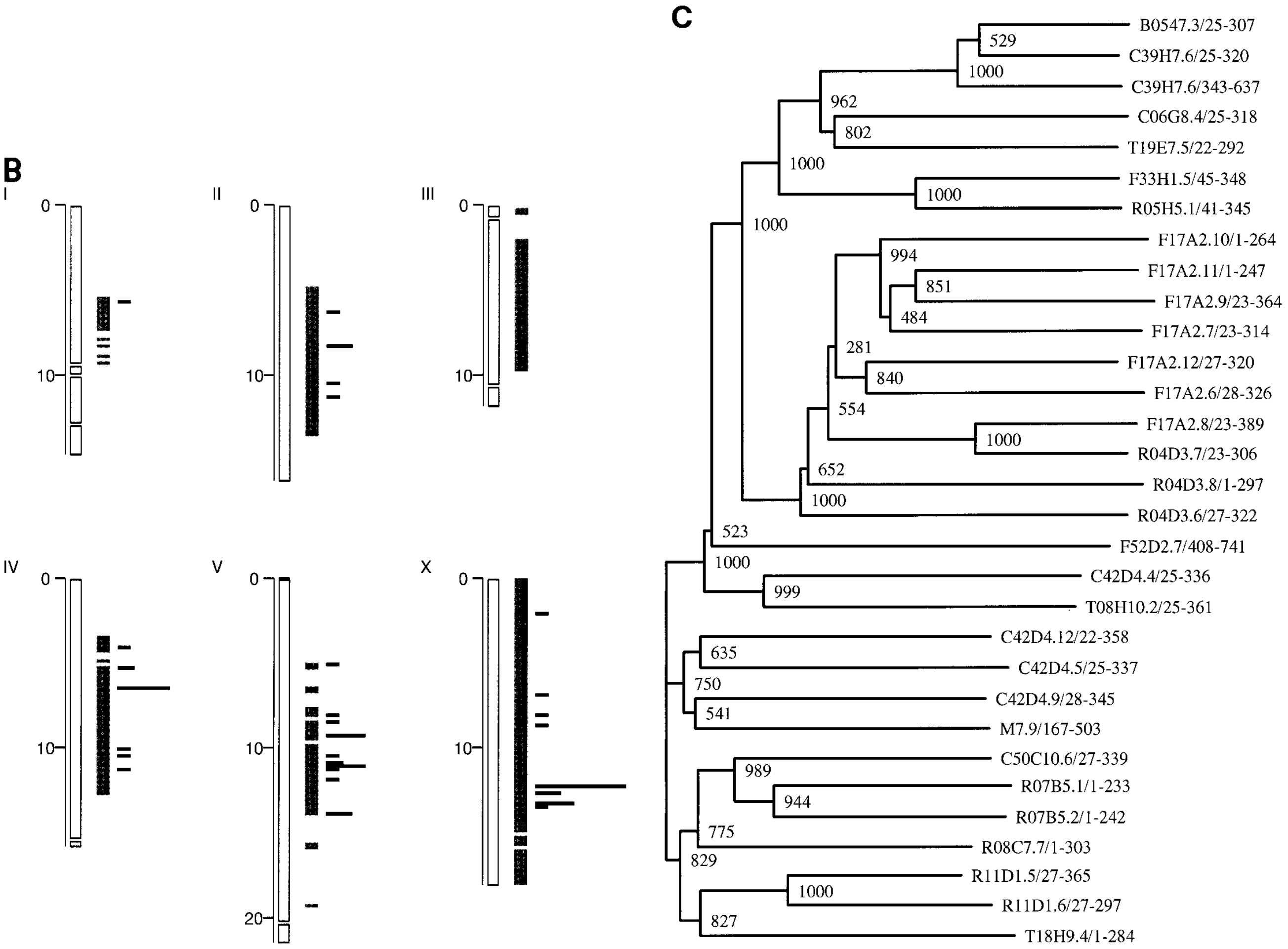


FIG. 7—*Continued*

to protein sequences and one that tries to resolve the problems but may suffer from other side effects.

To illustrate this, let us consider the simplest clustering method, "single linkage." The principle works as follows: All proteins are compared to one another, and all significant pairwise matches are stored. The proteins are then linked together in clumps by joining all proteins that have at least one match to one of the proteins in the group. This method has no protection against the joining of unrelated clusters by multidomain proteins or by spurious matches. We applied single linkage clustering to Wormpep 11. Using Blastp filtered by MSPcrunch with a relatively stringent cutoff (twilight zone between scores 40 and 80) resulted in a super-cluster containing about a third of all sequences. MSPcrunch effectively removes biased composition matches, but some spurious links will be accepted with these parameters. Raising the stringency to exclude all matches scoring below 90 eliminated essentially all spurious matches, but extensive joining of unrelated clusters still occurred due to multidomain proteins. The largest cluster contained 585 proteins, including such diverse protein families as protein kinases, phosphatases, proteases, protease inhibitors, transcription factors, and extracellular domains.

Methods exist to separate such clusters by manual inspection of a matching matrix (Watanabe and Otsuka, 1995), by match overlap analysis [the CLUSDOM program (Koonin *et al.*, 1996b)], or by minimal spanning tree analysis of segment pair regions (Hunter *et al.*, 1992, States *et al.*, 1993). We used the Domainer algorithm (Sonnhammer and Kahn, 1994), which explicitly takes domains into account. Domainer keeps a graph of segment information for each single-linkage cluster and uses this to split them where sequences diverge and at sequence ends. The main drawback of Domainer is that it is vulnerable to imperfections in its input of pairwise similarity data. Incomplete matching regions can cause Domainer to infer too many domain boundaries, resulting in fragmentation of real domains. Because of this overfragmentation, Domainer output often needs to be processed manually to produce true domain families. However, the core domains are usually of reasonable quality to use as a starting point, and the risk of merging unrelated families is small.

Domain-wise clustering of Wormpep. We can improve the Domainer clustering by using the previously found matches to Pfam-A families. By removing these matching segments, most of the large families, which are most prone to errors, are avoided, and correct domain boundaries are introduced. The procedure we used was to extract all sequence sections larger than 30 residues that were not covered in Pfam-A into separate entries. A protein with a Pfam-A domain in the center that has long flanking regions on either side will thus generate two entries. By doing this, Domainer will consider each section an independent sequence, and the boundary to the Pfam-A segment will be used as a real domain boundary.

Removal of the Pfam-A matching segments from Wormpep 11 left 8221 segments, which were clustered using Blastp and Domainer (see Materials and Methods). This yielded 1516 clusters between 2 and 58 members each, and 8602 singleton segments. The distribution of cluster sizes is shown in Fig. 4. To analyze what proportion of these are specific for *C. elegans* and other species in the phylum Nematoda, the consensus sequence of each cluster was searched against all known nucleotide sequences using Tblastn. Families with matches outside Nematoda are shown in black on Fig. 4; these account for 43% of the clusters. Most of the large-domain families found by Domainer appear to be specific to Nematoda, or at least detectable homologues have not yet been sequenced in other organisms. When Pfam-A matching segments were *not* removed from Wormpep, 80% of the clusters had matches outside Nematoda.

Apparently Nematode-Specific Protein Domain Families

The largest protein clusters that were apparently nematode specific were analyzed in further detail. To improve the quality of the alignments they were rebuilt from complete sequences. These alignments were searched against SwissProt and SwissProt-TREMBL using HMM methods (Eddy, 1996) as a second pass to look for matches to other organisms. Only families lacking clear homology outside Nematoda were considered. This way we have collected 10 apparently nematode-specific families, which are listed in Table 2. Hydrophobicity patterns, coiled coil predictions, and Prosite pattern matches were analyzed to give additional clues to the function.

Three of the families are probably G-protein-coupled receptors. Although the sequence similarity is weak, it is supported by alternating hydrophobic/hydrophilic regions typical of receptors, and there is also a characteristically conserved arginine at the end of the third predicted transmembrane helix. Some of the members are expressed in sensory neurons and are likely to function as olfactory receptors (Troemel *et al.*, 1995). The fact that G-protein-coupled receptor subfamilies do not find matches outside nematodes by standard sequence comparison methods is not surprising, since divergence rates of transmembrane proteins are typically higher than for globular proteins.

Three examples of these families are shown in Figs. 5–7. Family 2 (Fig. 5) has weak similarity to transthyretin (formerly called prealbumin), which transports thyroid hormones. The hydropathy plot of family 8 (Fig. 6) suggests it may have a transmembrane location, but there is no detectable specific similarity to known transmembrane sequence families. Family 9 (Fig. 7), the largest apparently nematode-specific family found, is likely to be a family of G-protein-coupled receptors. This is indicated by the hydrophobicity plot and the presence of a highly conserved arginine at the end of the third hydrophobic segment. The members in family 9 are not randomly distributed throughout the genome. As seen in Fig. 7B, large clusters are present on chromosomes V and X, while no member has been found in the extensively sequenced chromosome III. This suggests that these families arose by local gene duplication. There is also a strong correlation between the similarity and the distance between two members, which is illustrated by a tree in Fig. 7C. Members on the same cosmid are nearly always most similar to each other. The only exception in 10 cosmids with multiple family members is C42D4.4, which does not cluster with the three other members on cosmid C42D4.

Multiple alignments of these families are available by anonymous FTP at ftp.sanger.ac.uk in /pub/databases/wormpep/wormPfam.

Comparison of C. elegans to Other Genomes

One of the reasons for sequencing the genome of *C. elegans* was its importance as a model organism. Insights from nematode biology can often be extrapolated to human biology. For example, work on the cell death genes ced-3, ced-4, and ced-9 has helped uncover the apoptosis pathway regulating cell death in humans (Hengartner and Horvitz, 1994; Chinnaiyan *et al.*, 1997). Naturally there are differences, but many of the basic life-supporting functions involved in, e.g., energy metabolism, replication, gene expression, and signaling are conserved throughout all phyla. Since *C. elegans* is a multicellular animal with a broad range of tissues (nervous system, musculature, gut, etc.) it is hoped that many, if not most, human proteins will have a homologue in the worm. Many events during early development and differentiation, such as body patterning by the Hox cluster, are similar in the two organisms. To address the question of how much protein homology can be expected between human and *C. elegans*, Wormpep was compared with all human proteins in SwissProt.

It has been proposed that most protein domains that are present in two species belonging to different phyla are also found in many other phyla. In 1993, it was estimated that over 90% of these "ancient conserved domains" (ACRs) were already present as functionally characterized entries in the sequence databases (Green *et al.*, 1993). To reexamine the amount of conservation between organisms from different kingdoms, we have also compared the *C. elegans* proteins to the proteins in two completely sequenced genomes: the yeast *S. cerevisiae* and the eubacterium *H. influenzae.*

Pairwise comparison of proteins in H. sapiens, C. elegans, S. cerevisiae, and H. influenzae. All proteins

TABLE 3

Cross-Species Protein Comparison

Homo sapiens proteins in SwissProt 33	3475 (~5%)	
Homo sapiens proteins that match Wormpep 11	2077	60%
Homo sapiens proteins that match *Saccharomyces cerevisiae*	1432	41%
Homo sapiens proteins that match *Haemophilus influenzae*	323	9%
C. elegans proteins in Wormpep 11	7263 (~50%)	
C. elegans proteins that match *Homo sapiens*	2378	33%
C. elegans proteins that match *S. cerevisiae*	2146	30%
C. elegans proteins that match *Haemophilus influenzae*	454	6%
Saccharomyces cerevisiae proteins in SwissProt 33 and TREMBL[a]	6719 (~100%)	
Saccharomyces cerevisiae proteins that match *Homo sapiens*	1929	29%
Saccharomyces cerevisiae proteins that match *C. elegans*	2447	36%
Saccharomyces cerevisiae proteins that match *Haemophilus influenzae*	908	13%
Haemophilus influenzae proteins	1680 (100%)	
Haemophilus influenzae proteins that match *Homo sapiens*	282	17%
Haemophilus influenzae proteins that match *C. elegans*	340	20%
Haemophilus influenzae proteins that match *Saccharomyces cerevisiae*	489	29%

Note. The percentages within brackets in the second column indicate what fraction of the genome the set of proteins represents.

[a] Only yeast TREMBL entries that were nonidentical to SwissProt entries.

from each genome were compared to all proteins of the other genomes (see Materials and Methods). The results are listed in Table 3 and are summarised in Fig. 8. The animals *H. sapiens* and *C. elegans* had the highest level of similarity, with 60% of the human proteins matching *C. elegans.* In general, the organism with the smaller genome has a larger proportion of its genome matching, and the organism with the larger genome has a larger number of proteins matching. In terms of percentages, *H. influenzae* is most similar to *S. cerevisiae,* which in turn is most similar to *C. elegans,* which in turn is most similar to *H. sapiens.* This is in agreement with the phylogenetic tree and increasing complexity of these organisms.

Common proteins in subsets of C. elegans, H. influenzae, and S. cerevisiae. To investigate further to what extent protein families are shared among organisms from different kingdoms, we also looked for proteins that intersect these genomes. We excluded *H. sapiens* from this analysis, since only about 5% of the human proteins have been sequenced completely. Because proteins come in families, two proteins in one genome may match one in another. When calculating the match intersection of two or more genomes, we therefore obtain different numbers of matching proteins for each genome being considered. Table 4 lists the numbers separately for each participating genome; the lowest of the numbers in each intersection is given in the diagram in Fig. 9. The fact that *S. cerevisiae* in most cases contains more proteins than its counterparts is partly due to difficulty of obtaining a completely nonredundant set of *S. cerevisiae* proteins (see Materials and Methods).

Almost all of the proteins shared between all three organisms belong to families whose function has been characterized experimentally. Of the 301 *H. influenzae* proteins, 256 had functional annotation provided by TIGR. We analyzed the remaining 45 proteins, and 39 could be assigned a function with high confidence, leaving only 6 genes without a putative function (HI0090, HI0174, HI340, HI0701, HI0719, and HI1715). It will be interesting to find out what the function of these proteins is. Given that these proteins are conserved throughout so many phyla, they are likely to be of fundamental importance.

The largest contribution to the set of proteins shared by *C. elegans* and *S. cerevisiae* that are not present in *H. influenzae* is due to the eukaryotic protein kinases. A number of other protein families are also specific to eukaryotes, such as histones, tubulin, and many of the proteins involved in transcription, translation, and replication. Thirty-nine proteins were unique to *C. elegans* and *H. influenzae* (Table 5). Many of them are metabolic enzymes involved in biosynthesis, but a wide variety of cellular roles is represented. The proteins fall into 11 of 13 functional categories defined by Tatusov *et al.* (1996). We were surprised to note strong similarity between DNA polymerase I (HI0856) in *H. influenzae* and W03A3.2 in *C. elegans,* spanning the entire DNA polymerase domain, yet no significant similarity to a

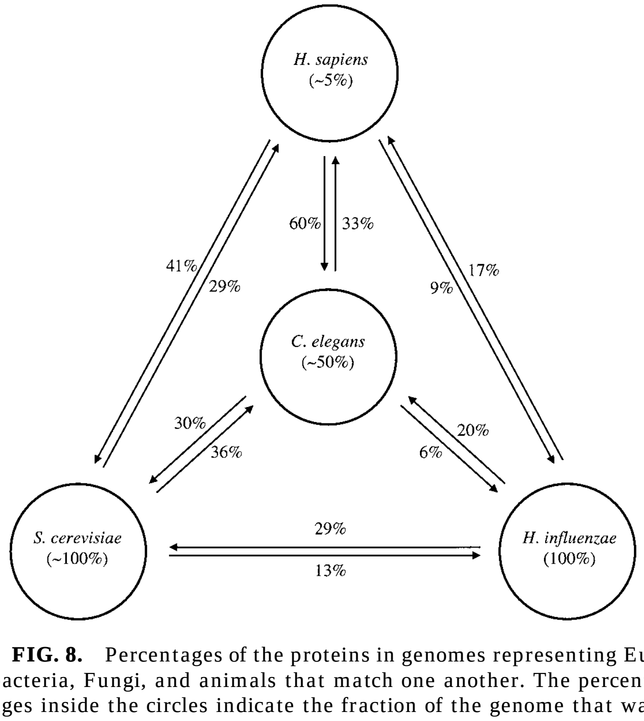


FIG. 8. Percentages of the proteins in genomes representing Eubacteria, Fungi, and animals that match one another. The percentages inside the circles indicate the fraction of the genome that was available for the analysis. See Table 5 for details.

TABLE 4

Common Subsets of Proteins Shared between Genomes from Organisms Representing Bacteria, Fungi, and Animalia

Organism combination	Proteins
Caenorhabditis elegans **not** (*Saccharomyces cerevisiae* **or** *Haemophilus influenzae*)	5049 (70%)
Saccharomyces cerevisiae **not** (*Caenorhabditis elegans* **or** *Haemophilus influenzae*)	3973 (59%)
Haemophilus influenzae **not** (*Caenorhabditis elegans* **or** *Saccharomyces cerevisiae*)	1135 (68%)
(*Caenorhabditis elegans* **and** *Saccharomyces cerevisiae*) **not** *Haemophilus influenzae*	1760 (24%), 1843 (27%)
(*Caenorhabditis elegans* **and** *Haemophilus influenzae*) **not** *Saccharomyces cerevisiae*	58 (1%), 39 (3%)
(*Saccharomyces cerevisiae* **and** *Haemophilus influenzae*) **not** *Caenorhabditis elegans*	299 (4%), 205 (12%)
Caenorhabditis elegans **and** *S. cerevisiae* **and** *Haemophilus influenzae*	396 (5%), 604 (9%), 301 (17%)

Note. In the overlap cases, where the numbers can be counted from either genome, they are listed in the same order as the species in the left column. Within brackets are the percentages of the genome that was counted from.

yeast protein was found. There is one yeast protein, the mitochondrial DNA polymerase Pol-γ, which is considered related to prokaryotic DNA polymerases (Braithwaite and Ito, 1993), but the sequence similarity is much weaker and confined to short motifs and was not significant with our method. A more detailed analysis of this case is presented elsewhere (Sonnhammer and Wootton, 1997). There are thus a number of instances in which patterns of sequence conservation clearly do not follow the groupings of the traditional phylogenetic tree of life.

When comparing three genomes with each other, a "bridging" situation that often occurs is when one genome contains a protein that is significantly similar to proteins from the two other genomes that do not show significant similarity to each other. We noted several cases of this in the *C. elegans*/*S. cerevisiae*/*H. influenzae* comparison. For example, *C. elegans* WP:F48E3.3 is similar to the killer toxin-resistance protein KRE5_YEAST (P22023) and to HI0259, but there is no discernible similarity between the *S. cerevisiae* and the *H. influenzae* proteins. In such cases, one could in principle infer homology indirectly. We have not pursued this strategy here, since most of the bridging cases require a much more thorough manual analysis to provide conclusive evidence of homology.

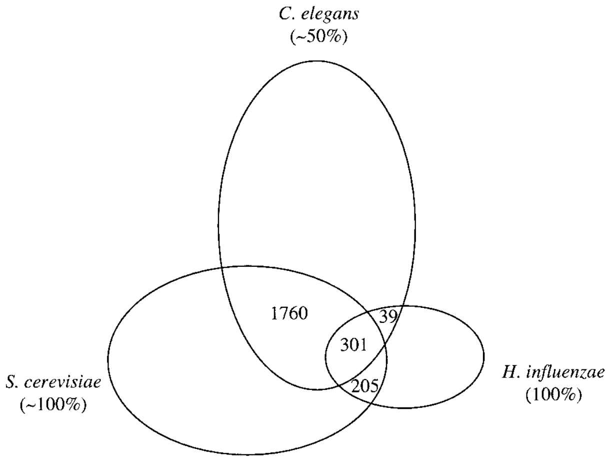


FIG. 9. Diagram of common proteins shared between three kingdoms. The numbers shown in intersecting areas are the lowest of the participating genomes. See Table 4 for details.

Human homologues in C. elegans. Nearly two-thirds of currently available human proteins have a homologue in 50% of *C. elegans* proteins. This figure can clearly not be doubled to predict the coverage when the whole *C. elegans* genome is available. The main reason for this is that most of the matching proteins belong to families of homologues. We expect that most protein families in *C. elegans* already have at least one representative in Wormpep 11 and that a majority of the human proteins that have a homologue in *C. elegans* already should have a match. To estimate the fraction of human proteins that will have a match to the entire *C. elegans* genome, we fitted a curve to a number of smaller sets of *C. elegans* proteins, as shown in Fig. 10. This curve suggests that approximately 70% of the human proteins in the set would match the entire *C. elegans* genome, which is only 10% more than the fraction that matches half of it.

Another factor that may be inflating the apparent proportion of human matches to *C. elegans* proteins is that the currently available human sequences may be biased toward widespread protein families found in model organisms. However, a similar result was obtained in a study of 70 positionally cloned human disease genes (Ahringer, 1997) (representing a sample without this bias), which found that 65% of these matched Wormpep 11.

As mentioned before, the number of isofunctional orthologues is significantly lower than the number of matching proteins. By "isofunctional orthologue" we mean homologues in different organisms that diverged due to speciation and still have the same biological role, as illustrated in Fig. 11. Since nonorthologous homologues may have diverged in function, they are often less useful for precise inference of biological information. We have estimated the number of isofunctional orthologues between the human and the *C. elegans* datasets by looking for homologues that are reciprocally the most similar pair of proteins, as seen from both genomes. This is usually the case for isofunctional orthologues. Counting every matching protein pair would overestimate this number significantly.

Further evidence for isofunctional orthology is that

TABLE 5

The 39 *Haemophilus influenzae* Proteins That Match *Caenorhabditis elegans* Proteins but Not *Saccharomyces cerevisiae*

H. influenzae ORF	Functional annotation
HI0019	*P*-methylase
HI0140	*N*-acetylglucosamine-6-phosphate deacetylase
HI0151	Membrane protease subunit
HI0152	Lipid synthesis enzyme (HetI family)
HI0211	Phosphatidylglycerophosphatase B
HI0244	Queuine tRNA-ribosyltransferase
HI0259	Lipopolysaccharide 1,2-glucosyltransferase
HI0280	Uridine phosphorylase
HI0392	Acyltransferase
HI0406	Acetyl-CoA carboxylase
HI0550	Glycosyl transferase
H10714	ATP-dependent Clp protease proteolytic subunit
HI0736	Sodium-dependent amino acid transporter
HI0765	Glycosyl transferase
HI0773	3-Oxoacyl CoA-transferase
HI0774	3-Oxoadipate CoA-transferase
HI0856	DNA polymerase I
HI0910	Mutator (AT-GC transversion) 8-oxo-dGTPase
HI1013	Sugar isomerase or lyase
HI1042	Methionine synthase
HI1075	Cytochrome d complex terminal oxidase
HI1115	Thioredoxin
HI1116	Deoxyribosephosphate aldolase
HI1219	Cytidylate kinase
HI1260	Acetyl-CoA carboxylase β-subunit
HI1324	Lon/Sms-related endopeptidase (no ATPase domain)
HI1362	NAD(P) transhydrogenase subunit
HI1363	NAD(P) NAD(P) transhydrogenase subunit
HI1441	Stringent starvation protein A, glutathione transferase
HI1448	Molybdopterin biosynthesis protein
HI1526	Cytidylyltransferase + sugar kinase
HI1545	Amino acid permease
HI1588	Formyltetrahydrofolate hydrolase
HI1646	Cytidylate kinase
HI1663	Glyoxalase
HI1675	Molybdenum cofactor biosynthesis protein
HI1676	Molybdenum cofactor biosynthesis protein
HI1690	GABA transporter
HI1705	Leucyl aminopeptidase

Note. The functional assignments were taken from the Tatusov and Koonin WWW server (Tatusov *et al.*, 1996) except in two cases (HI0019 and HI1663).

both proteins are about equally long and match over the entire length. Enforcing this criterion here is likely to lead to underestimation of the number of isofunctional orthologues, since the *C. elegans* proteins were predicted from genomic DNA, and may not always be complete, and because the extent of the match was estimated using Blastp, which may report only a part of a true match.

Of the 2077 human proteins that match *C. elegans*, 744 had reciprocally best partners. This number of proteins are thus likely to have isofunctional orthologues. Requiring that both proteins have to match each other over more than 80% of their lengths reduces the number to 257. Given that only about 5% of the human proteins were used in the analysis, many of the functionally identical orthologues may not have been sequenced yet. However, the human and *C. elegans* proteins that fulfil the stringent criteria mentioned above are likely to have very similar functions even if they are not mutually most similar when the whole genomes are compared. A further indication of orthology is that the two genes in question are more similar to each other than to homologues from phylogenetically more distantly related organisms (Tatusov *et al.*, 1996). For the human to *C. elegans* comparison, fungal, plant, or bacterial proteins could be used as outgroups. We only found one case in which the outgroup homologue was more similar than the human one. [The putative DNA helicase M03C11.2 is more similar to the yeast protein CHL1_YEAST (SwissProt P22516) than to the inferred human orthologue XPD_HUMAN (SwissProt P18074).] One reason for finding so few cases of this is that a large proportion of the putative orthologues is only found in animals.

The number of detectable isofunctional orthologue relationships should grow more linearly than the curve of homologues in Fig. 10. When the *C. elegans* genome is finished, we would expect a significant increase in isofunctional orthologue relationships compared to now, although probably not twice as many. A greater increase in functionally orthologous partners will come from sequencing the complete human genome. Given that *C. elegans* is estimated to contain no more than 15,000 genes, and only a third of the homologues above were deemed likely orthologues, it is not reasonable to expect more than 5000 eventual isofunctional orthologue pairs.

We refrained from performing the opposite extrapolation of what fraction of *C. elegans* proteins would match larger fractions of the human genome, since basing such an estimate on the currently available 5% of all human proteins would make the number at 100% highly unreliable.

DISCUSSION

This paper has provided a glimpse into what we can expect to learn from the complete genome sequence of a higher eukaryote. Some results were surprising while others were more or less expected. Molecular biology research before the genome projects had already indicated that many protein domains are conserved between distantly related organisms, while some appear to be unique to certain phylogenetic groups. With entire genome sequences such notions can be quantified, and detailed answers can be given about the degree of conservation and distribution of protein families in an organism. We are still in an early learning phase of how to interpret these patterns, but it is clear that sequencing and analyzing entire genomes will have a profound impact on biology and guide experimental research in new interesting directions.

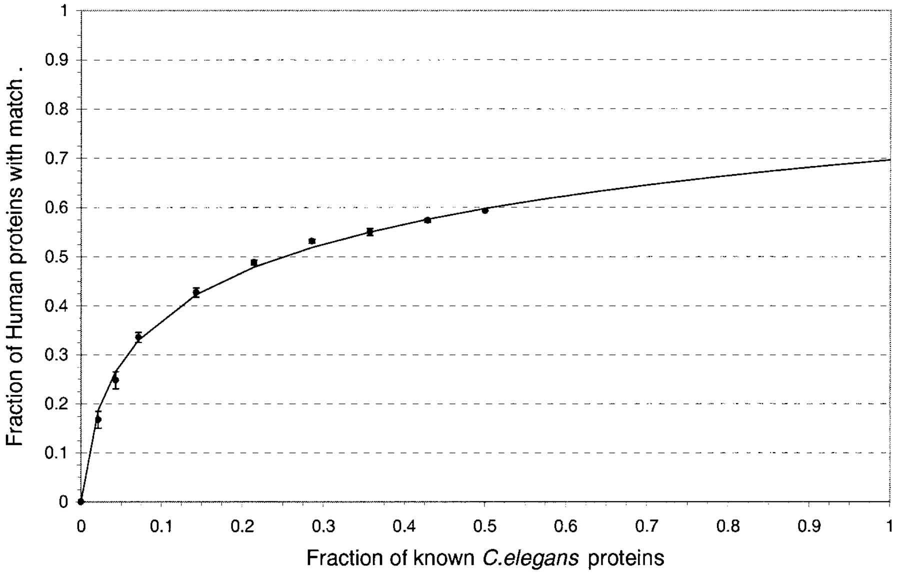


FIG. 10. Projection of the fraction of human proteins that match *C. elegans* proteins for different fractions of known *C. elegans* proteins. The data points below 50% were obtained as described for Fig. 3.

Perhaps one of the most striking results is the estimate that about 70% of the currently known human genes will have a homologue in the invertebrate *C. elegans.* This underlines the appropriateness and usefulness of studying this nematode, and we can expect that the unraveling of molecular biological phenomena in it will greatly assist the understanding of human biology. One should keep in mind, however, that the proportion of human homologues may decrease in the future as a less biased set of human genes is produced by genomic sequencing.

Another striking result is that most protein domains that are conserved in distantly related organisms have been biochemically characterized already. This is exemplified by the fact that of the 301 *H. influenzae* proteins also found in *C. elegans* and *S. cerevisiae,* only 7 had no functionally annotated homologues. This was also the case for only 1 of the 39 proteins found in *C. elegans* and *H. influenzae* but not in *S. cerevisiae.* This strongly supports the ACR theory (Green *et al.,* 1993), which was based on the observation that over 90% of newly found ACRs were already in the databases. Our analysis suggests that this figure is now at least 95%.

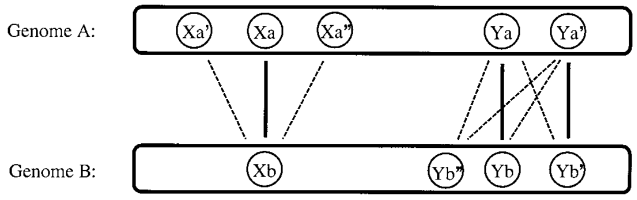


FIG. 11. When two genomes are compared, the number of matching proteins can be higher than the number of isofunctional orthologues. To illustrate this, consider the two genomes A and B, which are derived from an ancestor that contained one gene X and the paralogues Y and Y′. The solid lines indicate isofunctional orthologues. After the speciation divergence between A and B, there were two duplications in A of gene X, giving rise to Xa, Xa′, and Xa″. Although all these are technically orthologues of Xb, we distinguish Xa from the others as the isofunctional orthologue, on the basis that it maintained the original function while the others diverged. Establishing isofunctional orthology is not trivial; a number of methods are discussed in the text. Similarly, the gene duplication of Yb gives rise to more matches than isofunctional orthologues.

One of the most fundamental questions in bioinformatics is how much functional information can be inferred from a particular similarity. Obviously, the more sequence similarity between two proteins, the more likely they are to have similar functions. We have addressed this in the *C. elegans* to *H. sapiens* comparison by looking for putative isofunctional orthologues according to a number of criteria, and we found that it is likely to be true for 15–30% of the homologies. Nonorthologous homology, which often has a lower level of similarity, still allows many general features to be inferred, such as putative nucleotide binding moieties, protein–protein interaction domains, or catalytic activities. In such cases, the substrate(s) and the cellular role(s) can normally not be inferred from the homology. The scenario is more complicated if proteins in different organisms that perform identical functions, for instance a catalytic step in a metabolic pathway, have evolved from different ancestors. A number of such cases of "nonorthologous gene displacement" have recently been described (Koonin *et al.,* 1996a).

Homologous proteins (i.e., proteins that were derived from a common ancestor) often have similar sequences, because of the functional and structural constraints imposed on them. After long time spans, however, mutations accumulate, and the amino acid sequences may drift beyond the point of recognition. Performing the analysis on the basis of sequence similarity may therefore not necessarily give the ultimate answer. The methods based on comparing one sequence with another are at present probably close to as sensitive as they will ever get. It is possible to look further back in time by using multiple alignment methods, since then strongly conserved features stand out more prominently. This was exemplified here by the fact that many of the apparently nematode-specific families could be assigned a possible function when multiple alignments were studied. For example, as more members were gathered in families 4 and 5, it became increasingly clear that they were likely G-protein-coupled receptors. Some families, e.g., family 8, which still only has a small number of (very similar) members, continue to defy functional prediction. This may change in the future as more members are found.

Still, *C. elegans* and other organisms seem to contain a large number of unique families with few members. Are these truly unique protein families with novel folds? The answer must be sought with more sophisticated analysis methods than pure sequence comparison. Structural threading methods (e.g. Jones and Thornton, 1996; Moult, 1996) that fit a sequence to known structures to find the most likely fold may give an answer. However, it is not always clear what functional information can be transferred when two proteins with no sequence similarity share the same fold.

Our capability to recognize homologies was improved by searching a database of preassembled protein families such as Pfam, in addition to traditional single-sequence database searching. Although the number of proteins that changed status from "function unknown" to "putative function" was not enormous, a large number of novel and supportive domain classifications were found.

Even with the aid of Pfam, the overall rate of functional annotation in *C. elegans* is 43%, which is low in comparison to bacterial and yeast genomes, which often reach 70–80% annotation (Bork *et al.*, 1995; Casari *et al.*, 1996; Tatusov *et al.*, 1996). There are a number of possible reasons for this. *C. elegans* is the first metazoan genome to be sequenced completely, and a large number of protein families specific for metazoans may yet be undetected. It is worth noting that the first complete genome of an archaean, *Methanococcus jannaschii*, also had a relatively low annotation rate (Bult *et al.*, 1996). Further, relatively few nematode genes have been studied experimentally; many of the yeast and bacterial similarities were to previously studied yeast and bacterial proteins. In addition, some of the predicted *C. elegans* genes may not actually be real genes, and as mentioned before, there are cases where Blast similarity to annotated proteins is obvious, but the annotation of the *C. elegans* protein has not been updated. The percentage of annotation also depends on what is considered a significant annotation.

The clustering of *C. elegans* proteins and the distribution of the families that appear to be nematode-specific provide insights into the general mechanism of evolution of paralogues within a genome. It seems that chunks of DNA are duplicated and are preferentially inserted near the original. In many cases, the resulting duplicated gene will be rendered inactive by truncation or mutation, but what we observe today in the living organism is the result of the accumulation of a countless number of "lucky accidents."

The fact that 39 proteins occur in both *C. elegans* and *H. influenzae*, but not in yeast, suggests that yeast does not contain a complete set of basic eukaryotic proteins, but has lost some during evolution, possibly compensating for them with other proteins that can emulate their function. Gene phylogenies that do not correspond to the species phylogeny may also be caused by replacement with a gene that was horizontally transferred from one organism to another, in which case the sequences are said to be xenologous (Gray and Fitch, 1983). Some of the observations could also simply be caused by very different rates of genetic drift of these proteins in yeast, which made it impossible to recognize true homologues. A more thorough analysis would be necessary to establish which hypothesis is most likely for each individual case.

An aspect of protein function that is of vital importance to biology is how proteins interact with each other in the network of pathways that make up a living cell. That regulation and signal transduction is a major aspect of metazoan life is evident from the large number of protein kinases, receptors, and transcription factors found in *C. elegans*. An important step toward understanding molecular mechanisms at this level is therefore to make a complete inventory of all the known protein modules in this animal.

ACKNOWLEDGMENTS

We thank Sean Eddy, Daniel Lawson, Eugene Koonin, and Julie Ahringer, for comments on the manuscript, and the *C. elegans* sequencing project at the Sanger Centre and the Genome Sequencing Center, Washington University, St. Louis, for generating the DNA sequence data. The Sanger Centre is supported by the Wellcome Trust.

REFERENCES

Ahringer, J. (1997). Turn to the worm! *Curr. Opin. Genet. Dev.* **7:** 410–415.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *J. Mol. Biol.* **215:** 403–410.

Bairoch, A., and Apweiler, R. (1997). The SWISS-PROT protein sequence data bank and its supplement TREMBL. *Nucleic Acids Res.* **25:** 31–36.

Bairoch, A., Bucher, P., and Hofmann, K. (1997). The PROSITE database, its status in 1997. *Nucleic Acids Res.* **25:** 217–221.

Bisson, G., and Garreau, A. (1995). APIC: A generic interface for sequencing projects. *In* "ISMB-95: Proceedings Third International Conference on Intelligent Systems for Molecular Biology," pp. 57–65, AAAI Press, Menlo Park, CA.

Braithwaite, D. K., and Ito, J. (1993). Compilation, alignment, and phylogenetic relationships of DNA polymerases. *Nucleic Acids Res.* **21:** 787–802.

Bork, P., Ouzounis, C., Casari, G., Schneider, R., Sander, C., Dolan, M., Gilbert, W., and Gillevet, P. M. (1995). Exploring the *Mycoplasma capricolum* genome: A minimal cell reveals its physiology. *Mol. Microbiol.* **16:** 955–967.

Brenner, S. E., Hubbard, T., Murzin, A., and Chothia, C. (1995). Gene duplications in *H. influenzae. Nature* **378:** 140.

Bult, C. J., White, O., Olsen, G. J., Zhou, L., Fleischmann, R. D., Sutton, G. G., Blake, J. A., FitzGerald, L. M., Clayton, R. A., Gocayne, J. D., Kerlavage, A. R., Dougherty, B. A., Tomb, J.-F., Adams, M. D., Reich, C. I., Overbeek, R., Kirkness, E. F., Weinstock, K. G., Merrick, J. M., Glodek, A., Scott, J. L., Geoghagen, N. S., Weidman, J. F., Fuhrmann, J. L., Nguyen, D., Utterback, T. R., Kelley, J. M., Peterson, J. D., Sadow, P. W., Hanna, M. C., Cotton, M. D., Roberts, K. M., Hurst, M. A., Kaine, B. P., Borodovsky, M., Klenk, H.-P., Fraser, C. M., Smith, H. O., Woese, C. R., and Venter, J. C. (1996). Complete genome sequence of the methanogenic archaeon, *Methanococcus jannaschii. Science* **273:** 1058–1073.

Casari, G., De Daruvar, A., Sander, C., and Schneider, R. (1996). Bioinformatics and the discovery of gene function. *Trends Genet.* **12:** 244–245.

Chinnaiyan, A. M., O'Rourke, K., Lane, B. R., and Dixit, V. M. (1997). Interaction of CED-4 with CED-3 and CED-9: A molecular framework for cell death. *Science* **275:** 1122–1126.

Durbin, R., and Thierry-Mieg, J. (1996). *In* "The ACEDB Genomic Database," World Wide Web. [URL: ftp://ftp.sanger.ac.uk/pub/acedb]

Eddy, S. R. (1995). *In* "The HMMER Package," World Wide Web. [URL: http://genome.wustl.edu/eddy/hmmer.html#hmmer]

Eddy, S. R. (1996). Hidden Markov models. *Curr. Opin. Struct. Biol.* **6:** 361–365.

Gray, G. S., and Fitch, WM. (1983). Evolution of antibiotic resistance genes: The DNA sequence of a kanamycin resistance gene from Staphylococcus aureus. *Mol. Biol. Evol.* **1:** 57–66.

Green, P., Lipman, D. J., Hillier, L., Waterson, R., State, D., and Claverie, J.-M. (1993). Ancient conserved regions in new gene sequences and the protein databases. *Science* **259:** 1711–1716.

Gribskov, M., McLachlan, A. D., and Eisenberg, D. (1987). Profile analysis: Detection of distantly related proteins. *Proc. Natl. Acad. Sci. USA* **84:** 4355–4358.

Hengartner, M. O., and Horvitz, H. R. (1994). Programmed cell death in *Caenorhabditis elegans. Curr. Opin. Genet. Dev.* **4:** 581–586.

Henikoff, S., and Henikoff, J. G. (1991). Automated assembly of protein blocks for database searching. *Nucleic Acids Res.* **19:** 6565–6572.

Hillis, D. M., and Moritz, C. (1990). "Molecular Systematics," Sinauer, Sunderland, MA.

Hodgkin, J., Plasterk, R. H., and Waterston, R. H. (1995). The nematode *Caenorhabditis elegans* and its genome. *Science* **270:** 410–414.

Hunter, L., Harris, N., and States, D. J. (1992). Efficient classification of massive, unsegmented datastreams. *In* "International Machine Learning Workshop," pp. 224–232, San Mateo, CA.

Jones, D. T., and Thornton, J. M. (1996). Potential energy functions for threading. *Curr. Opin. Struct. Biol.* **6:** 210–216.

Koonin, E. V., Bork, P., and Sander, C. (1994). Yeast chromosome III: New gene functions. *EMBO J.* **13:** 493–503.

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

Koonin, E. V., Tatusov, E. V., and Rudd, K. E. (1996b). Protein sequence comparison at genome scale. *Methods Enzymol.* **266:** 295–322.

Krogh, A., Brown, M., Mian, I. S., Sjoelander, K., and Haussler, D. (1994). Hidden Markov models in computational biology: Applications to protein modelling. *J. Mol. Biol.* **235:** 1501–1531.

Lupas, A., van Dyke, M., and Stock, J. (1991). Predicting coiled coils from protein sequences. *Science* **252:** 1162–1164.

Maidak, B. L., Olsen, G. J., Larsen, N., Overbeek, R., McCaughey, M. J., and Woese, C. R. (1997). The RDP (Ribosomal Database Project). *Nucleic Acids Res.* **25:** 109–111.

Moult, J. (1996). The current state of the art in protein structure prediction. *Curr. Opin. Biotechnol.* **7:** 422–427.

Rice, P., Lopez, R., Doelz, R., and Leunissen, J. (1995). EGCG 8.0. *Embnet News* **2:** 5–7.

Romesburg, H. C. (1989). "Cluster Analysis for Researchers," Krieger, Melbourne, FL.

Rost, B., Casadio, R., Fariselli, P., and Sander, C. (1995). Transmembrane helices predicted at 95% accuracy. *Protein Sci.* **4:** 521–533.

Rubin, G. (1996). Around the genomes: The *Drosophila* Genome Project. *Genome Res.* **6:** 71–79.

Scharf, M., Schneider, R., Casari, G., Bork, P., Valencia, A., Ouzounis, C., and Sander, C. (1994). GeneQuiz: A workbench for sequence analysis. *In* "ISMB-94: Proceedings Second International Conference on Intelligent Systems for Molecular Biology," pp. 348–353, AAAI Press, Menlo Park, CA.

Sonnhammer, E. L. L., and Durbin, R. (1994). A workbench for large-scale sequence homology analysis. *Comput. Appl. Biosci.* **10:** 301–307.

Sonnhammer, E. L. L., and Durbin, R. (1996). A dot-matrix program with dynamic threshold control suited for genomic DNA and protein sequence analysis. *Gene* **167:** GC1–10.

Sonnhammer, E. L. L., Eddy, S. R., and Durbin, R. (1997). Pfam: A comprehensive database of protein domain families based on seed alignments. *Proteins* **28:** 405–420.

Sonnhammer, E. L. L., and Kahn, D. (1994). Modular arrangement of proteins as inferred from analysis of homology. *Protein Sci.* **3:** 482–492.

Sonnhammer, E. L. L., and Wootton, J. C. (1997). Widespread eukaryotic sequences, highly similar to bacterial DNA polymerase I, looking for functions. *Curr. Biol.* **7:** R463–R465.

States, D. J., Harris, N. L., and Hunter, L. (1993). Computationally efficient cluster representation in molecular megaclassification. *In* "ISMB-93: Proceedings First International Conference on Intelligent Systems for Molecular Biology," pp. 387–394, AAAI Press, Menlo Park, CA.

Tatusov, R. L., Mushegian, A. R., Bork, P., Brown, N., Hayes, W. S., Borodovsky, M., Rudd, K. E., and Koonin, E. V. (1996). Metabolism and evolution of *Haemophilus influenzae* deduced from a whole-genome comparison with *Escherichia coli. Curr. Biol.* **6:** 279–291.

Thompson, J. D., Higgins, D. G., and Gibson, T. J. (1994). CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* **22:** 4673–4680.

Troemel, E. R., Chou, J. H., Dwyer, N. D., Colbert, H. A., and Bargmann, C. I. (1995). Divergent seven transmembrane receptors are candidate chemosensory receptors in *C. elegans. Cell* **83:** 207–218.

Walker, R. D., and Koonin, E. V. (1997). SEALS: A system for easy analysis of lots of sequences. *In* "ISMB-97: Proceedings Fifth International Conference on Intelligent Systems for Molecular Biology," pp 333–339, AAAI Press, Menlo Park, CA.

Watanabe, H., and Otsuka. J. (1995). A comprehensive repesentation of extensive similarity linkage between large numbers of proteins. *Comput. Appl. Biosci.* **11:** 159–166.

Waterston, R., and Sulston, J. (1995). The genome of *Caenorhabditis elegans. Proc. Natl. Acad. Sci. USA* **92:** 10836–10840.

Wilson, R., Ainscough, R., Anderson, K., Baynes, C., Berks, M., Bonfield, J., Burton, J., Connell, M., Copsey, T., Cooper, J., Coulson, A., Craxton, M., Dear, S., Du, Z., Durbin, R., Favello, A., Fulton, L., Gardner, A., Green, P., Hawkins, T., Hillier, L., Jier, M., Johnston, L., Jones, M., Kershaw, J., Kirsten, J., Laisster, N., Latreille, P., Lightning, J., Lloyd, C., Mortimore, B., O'Callaghan, M., Parsons, J., Percy, C., Rifken, L., Roopra, A., Saunders, D., Shownkeen, R., Sims, M., Smaldon, N., Smith, A., Smith, M., Sonnhammer, E., Staden, R., Sulston, J., Thierry-Mieg, J., Thomas, K., Vaudin, M., Vaughan, K., Waterston, R., Watson, A., Weinstock, L., Wilkinson-Sproat, J., and Wohldman, P. (1994). 2.2 Mb of contiguous nucleotide sequence from chromosome III of *C. elegans. Nature* **368:** 32–38.

Zhang, J., Ostell, J., and Rudd, K. (1994). ChromoScope: A graphic interactive browser for *E. coli* data expressed in the NCBI data model. *In* "Proc. 27th Annual Hawaii International Conference on System Sciences," pp. 58–67, IEEE Computer Society Press.