Pacific Symposium on Biocomputing 5:30-41 (2000)

PROTEIN FOLDS IN THE WORM GENOME

M. GERSTEIN, J. LIN, H. HEGYI
Department of Molecular Biophysics & Biochemistry
266 Whitney Avenue, Yale University
PO Box 208114, New Haven, CT 06520, USA

We survey the protein folds in the worm genome, using pairwise and multiple-sequence comparison methods (i.e. FASTA and PSI-blast). Overall, we find that ~250 folds match ~8000 domains in ~4500 ORFs, about 32 matches per fold involving a quarter of the total worm ORFs. We compare the folds in the worm genome to those in other model organisms, in particular yeast and *E. coli*, and find that the worm shares more folds with the phylogenetically closer yeast than with *E. coli*. There appear to be 36 folds unique to the worm compared to these two model organisms, and many of these are obviously implicated in aspects of multicellularity. The most common fold in the worm genome is the immunoglobulin fold, and many of the common folds are repeated in various combinations and permutations in multidomain proteins. In addition, an approach is presented for the identification of "sure" and "marginal" membrane proteins. When applied to the worm genome, this reveals a much greater relative prevalence of proteins with seven transmembrane helices in comparison to the other completely sequenced genomes, which are not of metazoans. Combining these analyses with some other simple filters allows one to identify ORFs that potentially code for soluble proteins of unknown fold, which may be promising targets for experimental investigation in structural genomics. A regularly updated worm fold analysis will be available from bioinfo.mbb.yale.edu/genome/worm .

1 Introduction

The recent completion of the *C. elegans* genome provides an opportunity to study the occurrence of protein folds on a truly large scale [1]. Such structural-genomics studies complement genome analysis that focuses solely on sequence families [2, 3]. In our study we used an approach of pairwise sequence comparison methods combined with the PSI-Blast multiple sequence method [4, 5] and membrane protein prediction to study folds in the worm genome.

Our work follows on a number of recent related surveys. In terms of worm genome analysis, Chervitz *et al.* established orthologous relationships between about 20% of worm and 40% of yeast proteins, referring them to core biological processes common in the two organisms [2]. Copley *et al.* focused mostly on worm-specific signaling proteins, outlining the most abundant protein families with a broad range of signaling functions [6]. In terms of more general genome surveys, we have done a number of related analyses comparing various aspects of protein structure, such as secondary structural composition and fold usage, between several recently sequenced genomes [7, 8, 9, 10, 11, 12]. Similar studies also have been carried out by other investigators [13, 14, 15, 36].

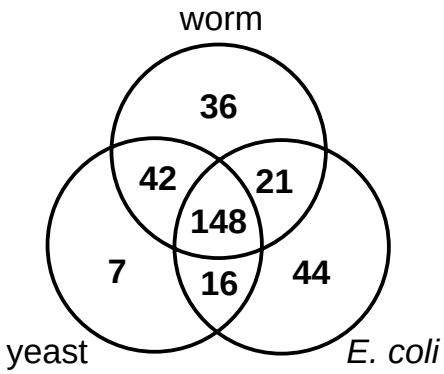
2 Fold Assignment Approach Used

The worm genome is substantially larger than any other genome sequenced. Consequently, we opted for an approach that allowed rapid automation, objective assessment, and *minimized* the number of false positives. We believe that the minimization of false positives is particularly important. We compared the structural domains classified in a recent version of scop (1.39) [16, 17] against the ~19000 predicted ORFs in worm using both PSI-blast and FASTA [5, 18, 19].

2.1 PSI-blast

We used the following parameters in our PSI-blast searches: an inclusion threshold (h) of 10^{-5}, the maximum number of iterations (j) of 10, and a final e-value cutoff of 1e-4. These parameters are somewhat stricter than those used in previous PSI-blast analyses -- e.g. our inclusion parameter is about 1/20 of that in Teichmann et al. [20] (h=0.0005, j=20). We monitored our parameter settings by seeing how many worm domains were assigned to two different protein folds (obviously an erroneous assignment) and made sure this number was virtually nil.

Figure 1, Fold Sharing between the Worm and other Model Organisms



The figure is a Venn diagram showing how 248 known protein folds are apportioned amongst the worm (*C. elegans*), yeast (*S. cerevisiae*), and *E. coli* genomes.

With our parameter choices we made up a set of fold templates based on the SCOP 1.39 domains to be used with PSI-blast, training them against nrdb90, applying a similar concept to that used in another recent survey [14].

2.2 FASTA

For the FASTA searches we used the usual e-value cutoff of .01 [21]. Although only 3113 ORFs were found to match with any of the SCOP domains, the vast majority of these matches were real positives, as reflected in the small number of the ORFs (only 15) that had matches with unrelated SCOP domains in the same region.

2.3 Web Presentation

The precise "fold counts" quoted below and in other recently published accounts are obviously contingent on the evolving state of the structural database and gene-prediction methods. A regularly updated worm fold analysis will, consequently, be available from bioinfo.mbb.yale.edu/genome/worm .

3 Analysis of Fold Assignments

3.1 Overall Coverage of the Genome: pairwise vs. multiple sequence methods

We find that 248 different protein folds match 7861 domains in 4586 worm proteins. On the fold level, this represents an approximately 32:1 level of duplication. The 248 folds are comprised of 304 structural superfamilies, representing an average 26:1 level of duplication.

Using simple pairwise comparisons with FASTA, the analogous numbers are considerably smaller, with only 138 folds and 160 superfamilies matching 4158 domains in 3033 worm proteins. Interestingly, the pairwise assignments give about the same level of duplication on the superfamily level (4148:160 ~ 26:1).

3.2 Fold Sharing between the Worm and other Model Organisms

As shown in Figure 1, it is informative to partition the folds in the worm into those shared and not shared with the complete genomes of other model organisms, in particular those of yeast and *E. coli* [22, 23]. Of the 314 total folds in the worm, yeast and E. coli genomes combined, 148 (48%) are shared by all three. Using only FASTA this number drops to ~110 (44%). Thus, the higher PSI-blast percentages indicate how more sensitive sequence comparison methods will tend to show a more encompassing list of primordial folds. These folds represent particularly ancient protein parts. Conversely, there are 36 unique worm-only folds, not present in yeast or *E. coli.* (But if we consider the other 17 completely sequenced microbial genomes

this number drops to 20.) These are probably "metazoan-only" folds, essential for carrying out worm-specific functions.

Table 1, Worm-only Folds

repre-sentative dom.	Name	scop class	scop fold num.	matches in worm genome
d1a68	Tetramerization domain of the shaker potassium channel	4	24	48
d1fim_	Tautomerase/MIF	4	41	2
d1kuh_	Zincin-like	4	52	60
d1dtp_	ADP-ribosylation	4	103	3
d1fid_	beta- and gamma-Fibrinogen Ct-dom.s	4	106	6
d1toh	Tyrosine hydroxylase catalytic and tetramerization domains	4	112	3
d2psi	Serpins	5	2	13
d1cii_	Toxin translocation dom.	6	1	1
d3ebx_	Snake toxin-like	7	6	1
d1bpi_	BPTI-like	7	7	138
d1ajj_	Ligand-binding domain of low -density lipoprotein receptor	7	11	116
d1krn_	Kringle modules	7	13	3
d1tur_	Ovomucoid/PCI-1 like inhibitors	7	14	23
d1ps2_	Trefoil	7	15	2
d1tgj_	Cystine-knot cytokines	7	16	5
d1hfi_	Complement control module	7	17	35
d2ech_	Blood coagulation inhibitor (disintegrin)	7	19	6
d1ata_	Ascaris trypsin inhibitor	7	22	38
d1exta1	Tumor necrosis factor (TNF) receptor	7	24	2

repre-sentative dom.	Name	scop class	scop fold num.	matches in worm genome
d1vii	Thermostable subdomain from chicken villin headpiece	1	14	2
d1pax_1	A domain of poly(ADP-ribose) polymerase	1	39	2
d1crka1	Creatine kinase, Nt-dom.	1	68	6
d1lla_1	Hemocyanin, N-terminal and middle domains	1	78	1
d1lbd	Ligand-binding domain of nuclear receptor	1	95	257
d1poc_	Phospholipase A2	1	103	2
d4aahb_	Non-globular all-alpha subunits of globular proteins	1	106	2
d1exg_	Common fold of diphtheria toxin/transcription factors/cytochrome f	2	2	22
d1npoa_	Neurophysin II	2	7	1
d1sfp_	Spermadhesin, CUB domain	2	18	84
d1aun_	Osmotin, thaumatin-like protein	2	19	6
d4fgf_	beta-Trefoil	2	31	8
d1hxn_	4-bladed beta-propeller	2	49	6
d1hcb_	Carbonic anhydrase	2	56	7
d1vmoa	beta-Prism I	2	59	23
d1cp3a	Caspase	3	11	5
d2tmda3	A nucleotide-binding domain	3	21	4

3.3 Breakdown of the 36 worm-specific folds

Of the 36 worm-specific folds listed in Table 1 as many as 25 (69%) are at least in part extracellular, as predicted by SignalP [24]. In contrast, only 18% of all the C. elegans proteins are believed to be secreted or partially extracellular [6]. The overwhelming predominance of the extracellular domains among the nematode-specific folds indicates the high impact multicellularity has had on the evolution of new folds, providing ways for the cells to communicate with one another. Almost all of the non-extracellular folds are also related in some ways to the multicellularity of the organism. For instance, both the tyrosine-hydroxylase or creatine-kinase folds play a role in energy transduction in tissues with high energy demands. Another example is the polymerase A domain (1.39), which is involved in differentiation.

In terms of broad structure-function classes, seven of the 36 folds are from the all-alpha class, and eight are from the all-beta class. The most highly represented SCOP structural class among the 36 is that of small proteins, with 11 belonging to this class, all of them extracellular. Only 9 folds of the 36 have enzymatic functions (5 of these are extracellular).

The most abundant of the worm-only folds is the all-alpha ligand-binding domain of nuclear receptor with ~250 domains in the worm genome. This fold contains a number of functionally divergent, versatile families of hormone receptors.

Table 2, Top-10 Worm Folds

best repre-sentative dom.	Name	scop class	scop fold num.	num. matches in worm genome (N)	frac. all worm dom. (F)	in EC?	in SC?
d1neu__	Immunoglobulin-like beta-sandwich	2	1	830	1.7%		
d1hev__	Knottins (Small inhibitors, toxins, lectins)	7	3	565	1.1%		
d1hcl__	Protein kinases (PK), catalytic core	5	1	472	0.9%		
d1lit__	C-type lectin-like	4	105	322	0.6%		
d1hcp__	Glucocorticoid receptor-like (DNA-bind dom.)	7	33	276	0.5%		
d1lbd__	Ligand-binding domain of nuclear receptor	1	95	257	0.5%		
d2bct__	alpha-alpha superhelix	1	91	247	0.5%		
d2adr__	Classic zinc finger, C2H2	7	31	239	0.5%		
d1gky__	P-loop Containing NTP Hydrolases	3	29	235	0.5%		
d1fxd__	like Ferrodoxin	4	34	207	0.4%		

F is estimated as N/T, where T is supposed to represent the total number of worm domains. This is estimated here as 50000 from the following relation: T = WM/P, where W is the average length of a worm ORF (450 residues), M is the total number of worm ORFs (19011), and P is the average length of a domain in the PDB (170 residues). The boxes in the last 2 columns are shaded if the fold occurs in *E. coli* or yeast (black, if more than 10 times.)

3.4 Worm vs. Yeast

As expected, the worm shares more folds with yeast than with the more phylogenetically distant *E. coli* (206 vs. 185). The analysis of the shared folds between worm and yeast illustrates particularly well how much more structure is conserved than sequence. Previously, based on straightforward pairwise sequence similarity, Chervitz et al. partitioned the ~19,000 proteins in the worm into those shared and not shared with yeast [2], with the latter group comprising 15446 "worm-only" proteins. However, these worm-only proteins contain 86 folds shared by yeast. The well-known TIM-barrel fold provides an excellent example of this fold sharing as it occurs often in each genome (>35 times). Some worm TIM-barrels (i.e. those in the "worm-only" set of Chervitz *et al.*) roughly correspond to yeast TIM-barrels

carrying out similar functions, suggesting they represent cases of marginal sequence similarity that could perhaps be detected by more sensitive comparison programs (e.g. worm C50B6.7 and yeast YBR299W or F01F1.12 and YKL060C). In contrast, other worm TIM-barrels are associated with functions unique to *C. elegans*, representing cases of a common scaffold acquiring worm-specific functions (e.g. 2K1058, a probable methylmalonyl-coA mutase precursor).

Table 3, The most frequent domain combinations in the worm genome

	SCOP Superfamilies		SCOP Superfamily Descriptions		Swissprot Repre-sentative
num.	dom. A	dom. B	domain A	domain B	
222	7.33.1	1.95.1	Glucocorticoid receptor (DNA-bind dom.)	Ligand-binding domain of nuclear receptor	ESTR_HUMAN
38	4.53.1	5.1.1	SH2 domain	Protein kinases (PK), catalytic core	FER_HUMAN
39	3.38.1	1.43.1	Thioredoxin-like	Glutathione S-transferases, C-term. dom.	SC1_OCTDO
27	4.105.1	2.18.1	C-type lectin-like	Spermadhesin, CUB domain	-
26	5.1.1	4.34.22	Protein kinases (PK), catalytic core	Adenylyl & guanylyl cyclase catalytic dom.	CYGF_HUMAN
22	5.7.1	1.24.6	Acyl-CoA dehydrogenase (flavoprotein) N-terminal & middle domains	Acyl-CoA dehydrogenase (flavoprotein) C-terminal domain	ACDM_HUMAN
19	3.29.1	1.56.1	P-loop containing NTP hydrolases	Transducin (A-subunit), insertion domain	GBAS_HUMAN
15	3.50.1	4.105.1	Integrin A (or I) domain	C-type lectin-like	-
14	2.1.1	2.1.2	Immunoglobulin	Fibronectin type III	AXO1_RAT
10	3.47.1	5.17.1	actin-like ATPase domain	Heat shock protein 70kD (HSP70) C-terminal, substrate-binding fragment	HS7C_HUMAN

3.5 The most common worm folds, in particular the Ig fold

The folds in the worm can be ranked in terms of their overall occurrence. The 10 most frequently occurring folds are listed in Table 2. The two most abundant folds, immunoglobulins (Ig) and knottins occur both in several extracellular proteins, often arranged in a large number of tandem repeats, or alternating with other types of domains. Both folds include several distinct, largely expanded superfamilies, as classified by SCOP. For instance, the Ig fold includes the immunoglobulin, cadherin, and fibronectin-3 superfamilies, and the knottin fold includes the EGF/Laminin superfamily.

Performing a detailed analysis of the 830 Ig domains, we found 452 representatives of the Ig superfamily (SCOP number: 2.1.1) in 77 ORFs, while other superfamilies of the Ig fold matched 378 domains in 166 ORFs. For the Ig superfamily we compared our results with the SMART collection of protein domains [25]. Despite the different approach used by SMART, their results largely agree with ours: 67 of the 77 ORFs were found by both methods, and most of the 10 additional ORFs were also identified by Pfam [3]. Further information about the Ig folds in the worm is available from bioinfo.mbb.yale.edu/genome/worm/ig.

Table 4, Commonly recombined worm folds

sfam#	Name	sing.	multi	tent. multi	repet-itive	combining domains
3.29.1	P-loop containing NTP hydrolases	38	58	78	15	ABCFGJLNOTVWX&$#.....
5.1.1	Protein kinases catalytic core	101	117	186	29	AGHIJLNOPRSTY&$@%.....
3.22.1	NAD(P)-binding Rossmann-folds	27	31	39	1	EKQXV#.....
7.3.9	Hairpin containg dom. of hep. GF	13	41	20	31	GHIUW%+@.....
3.7.1	Leucine-rich repeats	43	15	64	8	DGHKLORSXZ.....
2.24.2	SH3-domain	14	28	20	6	BCFGHILNPSYZ#..
1.91.3	Ankyrin repeat	33	18	38	31	LJMNSV$#^....
4.89.1	Glutathione syn. ATP-bind dom.	0	15	0	0	EQR.....
2.1.1	Immunoglobulins	2	39	11	20	CDIJUV#+^$@..
7.37.1	RING finger domain, C3HC4	48	10	89	5	GHMOSW#....

10 superfamilies combining with the greatest number of other superfamilies in multidomain proteins. Explanation of columns 3-6: number of times the superfamily was detected in single-domain, multidomain, tentatively multidomain and repetitive proteins, respectively. The last column indicates the superfamilies that each superfamily in column 1 combines with. Each letter or sign denotes one particular superfamily that occurs more than once in the table. Those superfamilies that occur only once are represented with a '.' sign - e.g. the 3.29.1 superfamily combines altogether with 26 other types of superfamilies in multidomain proteins, 16 occurring more than once and 10 only once in the table.

4 Multidomain proteins in the worm genome

Another possible way to analyze the fold occurrence in the worm genome is to partition the matching ORFs into three categories related to domain structure:

(1) single (only one matching domain in the ORF);

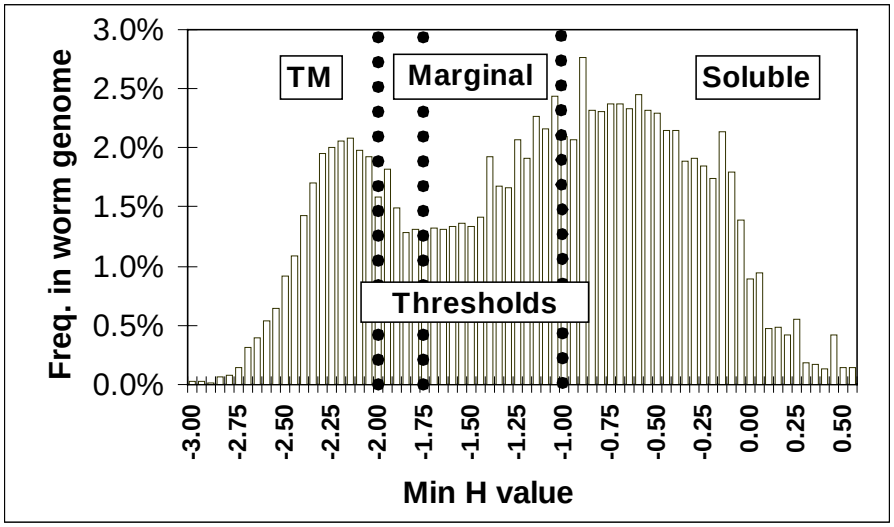
(ii) repetitive (the same type of domain matching an ORF several times); and

(iii) multidomain (not in category i or ii).

Of the 304 represented structural superfamilies, 224 were found in single, 191 in multidomain and 88 in repetitive proteins. Many of the superfamilies overlap among the three categories but several of them have a unique behavior occurring only in one of the three categories: 70 superfamilies are found only in multidomain, 88 only in single and 6 only in repetitive proteins.

Of the 4588 ORFs with structural matches, 3115 were found with a single domain match, 838 with multiple domains and 635 were repetitive. The highest number of matching domains were in the 5198-residue-long F15G9.4b protein, which contained a series of 46 Ig domains followed by a single EGF domain.

Figure 2, Criteria for Sure and Marginal Membrane Proteins



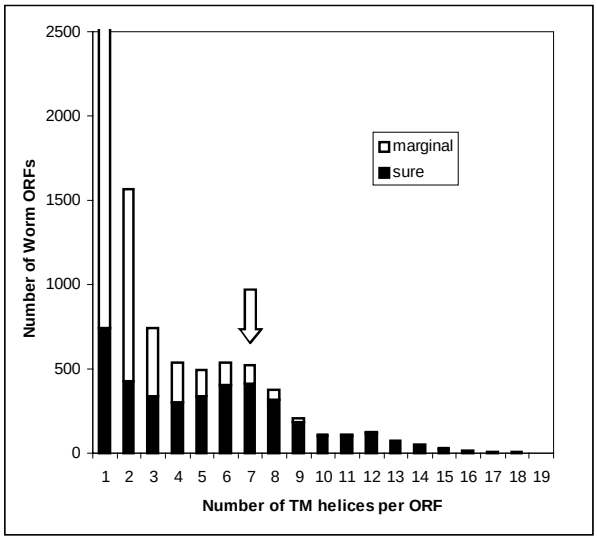
This shows a histogram of the MinH value (expressed in kcal/mole) for each worm ORF. MinH is the minimal value found in each sequence from running the GES-based TM identifier on it. The bimodal shape of its distribution suggests that a MinH value of -2 kcal/mole is a good discriminator between sure and marginal TM proteins.

In the multidomain proteins many of the domains combined preferentially with one another: Table 3 shows the ten most frequently occurring domain combinations (with their assigned SCOP superfamily numbers). Each combination is listed with a representative Swissprot protein with the same combination of domains. Interestingly, two frequently occurring combinations, the C-type lectin-like domain with CUB domain and the Integrin A/C-type domain with lectin-like domain, are not present in Swissprot at all, probably indicating worm proteins with novel functions.

Table 4 lists those superfamilies that combine with the highest number of other superfamilies. The top superfamily, the P-loop containing NTP-hydrolase was found to combine with as many as 26 other superfamilies in 58 multidomain proteins. Although the functions of proteins containing this superfamily are quite versatile, most of them bind to a nucleotide or to DNA. The second most favorably combining superfamily, the protein kinase catalytic core combines with 24 other superfamilies. In accordance with its name, it most frequently occurs in protein kinases, but also in several types of protein receptors. Of all the superfamilies in the list, this one occurs in the greatest number of proteins, not only in 117+186 multidomain ones but also in 101 instances as a single-domain protein. Interestingly, unlike the other superfamilies in the table, the glutathione-synthetase ATP-binding-domain

superfamily occurs only in multidomain proteins, combining with as many as 13 other superfamilies. However, only three of the combining superfamilies occur more than once in the table, indicating the relatively isolated nature of this superfamily.

Figure 3, Membrane Proteins in the Worm Genome

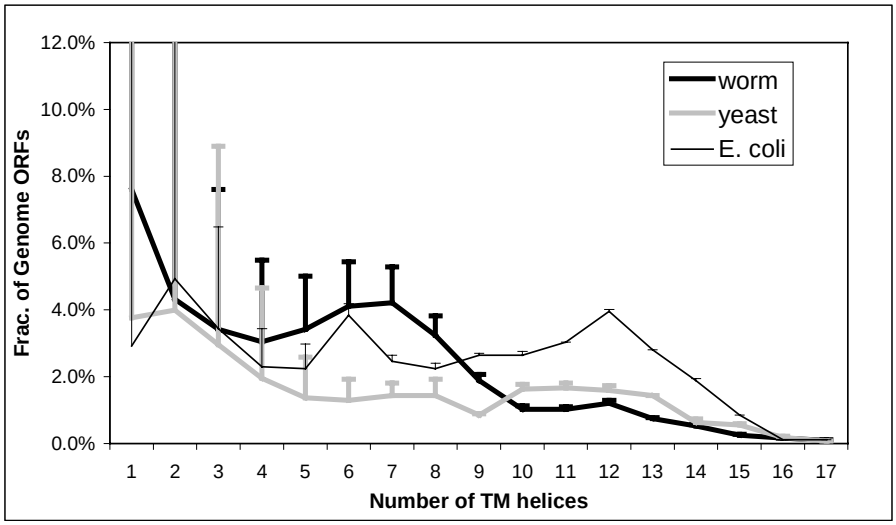


The figure shows the number of worm ORFs with a given number of predicted TM helices. The predictions are divided into sure and marginal. (See figure 2 for explanation.)

5 Membrane proteins in the worm genome

There have also been many surveys of the occurrence of membrane proteins in genome sequences [10, 11, 26, 27, 28, 29, 30, 31, 32]. The overall number of membrane proteins found depends somewhat on the prediction method and threshold used. Nevertheless, there seems to be a broad agreement that part or all of 20-30% of the proteins in microbial genomes are membrane proteins, with yeast having a slightly larger fraction.

Figure 4, Comparison of membrane proteins in the worm, yeast, and *E. coli*



The curves indicate the number of "sure" membrane proteins in three model organisms. Note the peaks at 6 and 12 TM segments for *E. coli* and 7 for the worm. The number of "marginal" membrane proteins is indicated by the bars extending upwards from the curves.

In our transmembrane (TM) identification strategy for the worm genome, TM segments were identified using the GES hydrophobicity scale [33]. The values from the scale for amino acids in a window of size 20 (the typical size of a transmembrane helix) were averaged and then compared against a cutoff of -1 kcal/mole. A value under this cutoff was taken to indicate the existence of a transmembrane helix. Initial hydrophobic stretches corresponding to signal sequences for membrane insertion were excluded. (These have the pattern of a charged residue within the first seven, followed by a stretch of 14 with an average hydrophobicity under the cutoff.) These parameters have been used, tested, and refined on surveys of membrane protein in genomes [29, 32, 34]. "Sure" membrane proteins had at least one TM-segment with an average hydrophobicity less than -2 kcal/mole. The rationale for this "MinH criteria" is the bimodal distribution of scores in figure 2. This is a similar approach to Boyd & Beckwith's MaxH criteria [30] and also the approach of Klein & Delisi [35]. "Marginal" membrane proteins had GES-identified TM-helices but did not fulfill the MinH criteria.

Our results are shown in figure 3 and 4. Overall, we find that about 21% (3964) of the ORFs in the worm genome are "sure" membrane proteins. Another 31% (5849) have more marginal membrane protein annotation. This seems to indicate

that the worm has a larger fraction of its genome devoted to membrane proteins than the other genomes sequenced so far, none of which are of metazoans [8]. Also notable, it appears that in the worm the distribution of membrane proteins in relation to their number of TM-helices has a peak around seven (see arrow in figure 3). This peak has not been found in previous surveys of membrane proteins in microbial genomes and probably results from the many 7-TM proteins that are needed for intercellular communication (see close-up in figure 4).

6 Conclusion

We surveyed the protein folds in the worm genome, using pairwise and multiple-sequence comparison methods (i.e. PSI-blast), and found that ~250 folds match ~8000 domains in ~4500 ORFs. We compared the folds in the worm genome to those in other model organisms, in particular yeast and E. coli, and found that worm shares more folds with yeast than with E. coli. There appear to be 36 folds unique to the worm compared to these two genomes, and many of these are obviously implicated in multicellularity. The most common fold in the worm genome is the Ig fold and many of the common folds occur in various repetitive combinations in multidomain proteins. Membrane protein analysis of the worm reveals a much greater relative prevalence of 7-TM proteins in comparison to the other completely sequenced genomes, which are not of metazoans.

Acknowledgments

We thank Amar Drawid for reading the text and the Keck and Donaghue Foundations for financial support.

References

1. Consortium, T.C.e.S. *Science* **282**, 2012-8 (1998).
2. Chervitz, S.A., *et al. Science* **282**, 2022-8 (1998).
3. Sonnhammer, E.L., Eddy, S.R., Birney, E., Bateman, A. & Durbin, R. *Nucleic Acids Res* **26**, 320-2 (1998).
4. Pearson, W.R. *Meth. Enz.* **266**, 227-259 (1996).
5. Altschul, S.F., *et al. Nucleic Acids Res* **25**, 3389-402 (1997).
6. Copley, R.R., Schultz, J., Ponting, C.P. & Bork, P. *Curr Opin Struct Biol* **9**, 408-15 (1999).
7. Hegyi, H. & Gerstein, M. *J Mol Biol* **288**, 147-64 (1999).
8. Gerstein, M. & Hegyi, H. *FEMS Microbiology Reviews* **22**, 277-304 (1998).
9. Gerstein, M. *Folding & Design* **3**, 497-512 (1998).
10. Gerstein, M. *Proteins* **33**, 518-534 (1998).

11. Gerstein, M. *J. Mol. Biol.* **274**, 562-576 (1997).
12. Teichmann, S, Chothia, C & Gerstein, M (1999). *Curr. Opin. Struct. Biol.* **9**: 390-399..
13. Frishman, D. & Mewes, H.-W. *Nature Struct. Biol.* **4**, 626-628 (1997).
14. Wolf, Y.I., Brenner, S.E., Bash, P.A. & Koonin, E.V. *Genome Res* **9**, 17-26 (1999).
15. Fetrow, J.S., Godzik, A. & Skolnick, J. *J Mol Biol* **282**, 703-11 (1998).
16. Murzin, A., Brenner, S.E., Hubbard, T. & Chothia, C. *J. Mol. Biol.* **247**, 536-540 (1995).
17. Hubbard, T.J.P., Murzin, A.G., Brenner, S.E. & Chothia, C. *Nucleic Acids Res* **25**, 236-9 (1997).
18. Pearson, W.R. *J Mol Biol* **276**, 71-84 (1998).
19. Lipman, D.J. & Pearson, W.R. *Science* **227**, 1435-1441 (1985).
20. Teichmann, S., Park, J. & Chothia, C. *Proc. Natl. Acad. Sci.* **95**, 14658-63 (1998).
21. Brenner, S., Chothia, C. & Hubbard, T. *Proc. Natl. Acad. Sci. USA* **95**, 6073-6078 (1998).
22. Goffeau, A., Aert, R., Agostini-Carbone, M.L., Ahmed, A. & 632-names-in-total. *Nature* **387(Supp)**, 5-105 (1997).
23. Blattner, F.R., *et al. Science* **277**, 1453-74 (1997).
24. Nielsen, H., Brunak, S. & von Heijne, G. *Protein Eng* **12**, 3-9 (1999).
25. Ponting, C.P., Schultz, J., Milpetz, F. & Bork, P. *Nucleic Acids Res* **27**, 229-32 (1999).
26. Goffeau, A., Slonimski, P., Nakai, K. & Risler, J.L. *Yeast* **9**, 691-702 (1993).
27. Rost, B., Fariselli, P., Casadio, R. & Sander, C. *Prot. Sci.* **4**, 521-533 (1995).
28. Rost, B. *Meth. Enz.* **266**, 525-539 (1996).
29. Arkin, I., Brunger, A. & Engelman, D. *Proteins* **28**, 465-466 (1997).
30. Boyd, D., Schierle, C. & Beckwith, J. *Prot. Sci.* **7**, 201-205 (1998).
31. Jones, D.T. *FEBS Lett* **423**, 281-5 (1998).
32. Wallin, E. & von Heijne, G. *Protein Sci* **7**, 1029-38 (1998).
33. Engelman, D.M., Steitz, T.A. & Goldman, A. *Annual Review of Biophysics & Biophysical Chemistry* **15**, 321-53 (1986).
34. Tomb, J.-F., *et al. Nature* **388**, 539-547 (1997).
35. Klein, P., Kanehisa, M. & DeLisi, C. *Biochim. Biophys. Acta* **815**, 468-76 (1985).
36. Mewes, H.W., *et al. Nature* **387**, 7-65 (1997).