Digging for Dead Genes: An Analysis of the Characteristics and Distribution of the Pseudogene Population in the Ribbon Worm Genome

Paul M. Harrison, Nathaniel Echols and Mark B. Gerstein *

Dept. of Molecular Biophysics & Biochemistry,

Yale University,

260 Whitney Ave.,

P.O. Box 208114,

New Haven, CT 06511-8114,

U.S.A.

**corresponding author*

Telephone: (203) 432-6105, (203) 432-5065

Fax: (360) 838 7861

E_mail: Mark.Gerstein@yale.edu

Submitted to Nucleic Acids Research

August 17th, 2000

Abstract

Pseudogenes are non-functioning copies of genes in genomic DNA, which may either result from reverse transcription from a messenger RNA transcript (termed processed pseudogenes) or from gene duplication and subsequent disablement (non-processed pseudogenes). As pseudogenes are apparently 'dead', they usually have a variety of disablements (e.g. insertions, deletions, frameshifts and truncations) relative to their functioning homologues. We have derived an initial estimate of the size, distribution and characteristics of the pseudogene population in the ribbon worm (*Caenorhabditis elegans)* genome, performing a survey in 'molecular archaeology'. Based on the estimated 18,576 proteins in the worm (*i.e.,* in the Wormpep18 database), we have found 3,814 corresponding pseudogenes and pseudogenic fragments, about one for every 5 genes. Few of these appear to be processed. Details of our pseudogene assignments are available from http://bioinfo.mbb.yale.edu/genome/worm/intergenic. The population of pseudogenes differs significantly from that of genes in a number of respects: (i) Pseudogenes are distributed unevenly across the genome relative to genes, with a disproportionate number on chromosome IV; (ii) The density of pseudogenes is higher on the arms of the chromosomes, unlike the density of genes; (iii) The amino-acid composition of pseudogenes is midway between that of genes and (translations of) random intergenic DNA, with the largest differences for Phe, Ala, Asp and Glu; And (iv) the most common protein folds and families differ between genes and pseudogenes -- whereas the most common protein fold found in the worm proteome is the immunoglobulin fold, the most common 'pseudofold' is that of phosphoglycerate mutase, which is only moderately abundant in the proteome. In addition, the size of a gene family bears no relationship to the size of its corresponding pseudogene

complement, indicating a highly dynamic genome. In fact, there are some striking examples of almost or completely extinct gene families associated with large populations of pseudogenes. For example, one worm chemoreceptor gene (D1022.6) has an extensive array of related ‘dead’ genes, but no obvious paralogs, thus perhaps representing a nearly extinct family, and another pseudogene family is only homologous to a non-worm protein, yeast gene YJL007C.

Keywords: pseudogene, molecular evolution, protein folds, proteome, *C. elegans*

Introduction

Over the course of evolution, genes duplicate in the genome, gradually accumulating mutations that may lead to the acquisition of new functions, or to the modification of existing functions. However, some duplications of genes acquire deleterious mutations that disable them so that they can no longer be translated into a functional protein. The disablement may occur at either or both the transcription and translation levels. These copies of genes are called non-processed pseudogenes. Pseudogenes may also arise by a process of retrotransposition, where a messenger RNA transcript is reverse transcribed and re-integrated into the genome [1-3]. These are termed processed pseudogenes or retropseudogenes and occur in a variety of plants and animals.

Some pseudogenes are evidently transcribed. A possible case of a ‘functioning’ pseudogene transcript has been described recently for neural nitric oxide synthase in the snail *Lymnaea stagnalis* [4]. Here, the pseudogene has a segment that is the inverse complement of the normal gene, and interferes through RNA duplex formation with the expression of nitric oxide synthase [4]. Interestingly, the expression of pseudogene transcripts can vary markedly with the expression of the transcripts of their paralogs. For example, for the 5-HT7 receptor, transcripts of a pseudogene can be detected in various tissues where transcripts for the corresponding functioning gene are absent [5]. Pseudogene transcripts can have raised expression in tumour cells, *e.g.* in a laryngeal squamous cell carcinoma [6] or in glioblastoma [7].

Pseudogenes are important sequences in the study of molecular evolution. They generally acquire mutations, insertions and deletions without any apparent evolutionary pressures, although in *Drosophila* for example, many putative pseudogenes appear to have patterns of mutation that are inconsistent with a lack of functional constraints [8-10]. Pseudogenes have been studied to derive underlying rates of nucleotide substitution [11-13] and insertion/deletion rates in genomic DNA [14, 15]. Averof, *et al.* [13] used eta-globin pseudogenes to show that double-nucleotide substitutions occur more often than would be expected from independent single-nucleotide substitutions. Gu and Li [14] noted that the pattern of insertions and deletions in processed pseudogenes implies that a logarithmic gap penalty dependence on gap size in sequence alignment is more appropriate than the more commonly used linear dependence. Ophir and Graur [15] performed a survey of processed pseudogenes in humans and murids and found evidence that gene truncations, insertions and deletions each occur by different mechanisms. Pseudogenes are also useful in determining rates of genomic DNA loss for an organism: a smaller complement of pseudogenes in a genome implies a greater net loss of genomic DNA [10, 16]. Petrov, *et al.* [16] recently demonstrated experimentally, using dead copies of retrotransposons as ‘pseudogene surrogates’, that the rates of DNA loss in *Drosophila* and the cricket *Laupala* are key determinants of genome size. In certain circumstances, pseudogenes can be conserved by a process of gene conversion, such as for immunoglobulin V_H pseudogenes in the chicken [17]. Goncalves, *et al.*

[18] surveyed human retropseudogenes and found that genes with a high number of retropseudogene copies tend to be widely expressed, highly conserved and low in GC content.

With the complete genomes of more than 30 prokaryotes and 4 eukaryotes (including the *Caenorhabditis elegans* genome [19]) now published, we have the opportunity to investigate pseudogenes on the genomic level. Surveys have recently been performed on the genes and pseudogenes of families of G-protein coupled receptors [20, 21]. We have conducted a global survey of the population of pseudogenes in the *Caenorhabditis elegans* genome. Our survey highlights some surprising characteristics of the pseudogene population, such as increased density of pseudogenes on the arms of the chromosomes and the distinct composition of the pseudogene population in terms of protein folds from that of the proteome. In a sense, our survey is a form of 'molecular archaeology', focussing on the characteristics of the 'dead' genes that can be uncovered in a genome. We see it as logically following upon a number of global surveys of the characteristics of the 'living' protein population in the newly sequenced genomes [22-24].

Results and Discussion

General definitions: G, Ψ G, and related terms

Given the gene population of the ribbon worm genome, what is the size and distribution of the corresponding pseudogene population? We have defined several populations of genes and pseudogenes in the present analysis (Table 1). The total population of confirmed and predicted protein-encoding genes (denoted *G*) is taken from the Wormpep18 database. The set of genes with at least one verifying EST was compiled (G_E). G_E was expanded by including all of the paralogs of G_E proteins to give the G_P set. Singleton genes are those genes that do not have an obvious paralog. The estimated population of pseudogenes that correspond to G is denoted as ΨG. In general, the symbol Ψ before any gene name or gene population name denotes a pseudogene or pseudogene population that corresponds to that gene or gene population. The use of the term pseudogene does not imply any attempt at parsing the exon structure, but refers loosely to any pseudogenes or pseudogenic fragments detected by homology matching and the occurrence of a simple disablement (whether a premature stop codon or a frameshift). Our survey is thus an initial estimate of ΨG that is extrapolated from the existing population of confirmed and predicted genes.

Estimated size of pseudogene population

The pseudogene population (denoted ΨG) arising from the decay of protein-coding genes in the ribbon worm is estimated to comprise 3,814 sequences. This corresponds to 21% of the total number of genes (Table 1). This is only an initial estimate of the pseudogene complement, that may be examined for broad trends and characteristics. However, there are a number of obvious factors that may affect the size of ΨG here. Firstly, dead copies of transposable elements would lead to an over-estimate of ΨG. However, these may be considered validly as pseudogenic fragments, and have been used as such in studies of DNA loss in *Drosophila* [10, 16]. Nonetheless, we do not find any abundant patterns of multiple protein-homology hits in the genomic DNA that would be indicative of a major unknown transposable element (see section below). Only ~5% of our total potential pseudogene matches are deleted because of matches to known transposable element proteins (see *Methods*). Secondly, the size of ΨG here may be an underestimate as we do not include pseudogenes that *only* have the less obvious coding disablements, such as damaged splicing signals. Thirdly, some of the pseudogenes may arise because of sequencing errors (and so should be annotated as genes). However, the reported overall error rate in sequencing is low (<1 in 10,000 bases) [25]. Fourthly, some pseudogenes may be fragments of two separate pseudogenes;

however, this problem is minimized by merging some pseudogene matches along the genomic DNA, with a procedure described below in *Method*s.

Pseudogene subpopulations

When only EST-matched genes are considered, ΨG_E corresponds to 13% of G_E (997 predicted pseudogenes) (Table 1). This EST-matched proportion can be considered a lower bound of more confidently predicted pseudogenes. (Intermediate between these, there are 2,729 predicted pseudogenes that correspond to a gene with an EST match or that are paralogous to a gene with an EST match, ΨG_P.) This may indicate that pseudogenes occur with greater frequency for more lowly expressed genes. Conversely, highly expressed genes may be less likely to have dead gene copies or fragments. Interestingly, singleton genes have a smaller relative population of pseudogenes (corresponding to 11% of the total number of singleton genes) yet constitute 32% of the gene population.

Intronic pseudogenes are pseudogenes that are contained completely within a single intron. A substantial fraction of ΨG is intronic (30%) (Table 1). Interestingly, there is no preference for sense or antisense alignment for an intronic pseudogene relative to the exons of the surrounding gene (51% are antisense). This indicates that the existence of pseudogenes in an intron has no relation to the transcription and splicing of a gene.

A key consideration is the proportion of ΨG that is predicted to be processed pseudogenes. Processed pseudogenes are derived originally from messenger RNA transcripts that have been reverse-transcribed and re-integrated into the genome. They have the following features: (1) they lack the introns of the gene from which they are derived; (2) they tend to have a characteristic polyadenine tail; (3) they lack the promoter structure of the gene from which they are derived; (4)

they have short direct repeats (about 9-15 base pairs) at their N- and C-termini [2]. We could not find any mention of processed pseudogenes in the ribbon worm in the scientific literature. We estimated the proportion of processed pseudogenes in ΨG using a simple heuristic that involved looking for stretches of coding sequence that could not be in the pseudogene without processing (which we have termed 'exon seams') and also for evidence of a polyadenine tail (see *Methods*). According to the exon seams identification, there appear to be few pseudogenes that result from processing in ΨG (totalling 299, 8%). We could not find any obvious subpopulation of pseudogenes with an elevated adenine content 3' to their homology segment that would indicate a polyadenine tail. The size of the estimated population of processed pseudogenes here contrasts substantially with the human genome, where about 80% of the pseudogenes are predicted to be processed [26].

Chromosomal distribution of pseudogenes

We mapped the positions of pseudogenes and genes along each of the six ribbon worm chromosomes. Pseudogenes are markedly more abundant nearer the ends or 'arms' of the chromosomes (Figure 1). When the distributions for the individual chromosomes are merged, we find that 53% of the pseudogenes are in the first and last 3 megabases of the chromosomes, compared to only 30% of the genes. It was previously noted [19] that the proportion of genes with similarities to other organisms tends to be lower on the chromosomal arms. The pseudogene distribution along chromosomes correlates with this observation and supports the idea of more rapidly evolving genomic DNA towards the ends of the chromosomes [19]. The same trend for increased occurrence of pseudogenes is observed for the G_E and G_P subsets, with 57% and 58% of the pseudogenes in the first and last 3Mb of genomic DNA respectively. This trend is depicted for the G_E set in Figure 1.

The distribution of pseudogenes between the chromosomes is also uneven (Figure 1 legend). For each chromosome, we calculated the proportion of 'dead' genes (equal to $|\Psi G_n| / [|\Psi G_n| + |G_n|]$, where $|G_n|$ is the size of the gene population G_n for chromosome *n* and $|\Psi G_n|$ is the number of pseudogenes). Chromosome IV appears to be the most 'dead', chromosome II the least (Figure 1 legend). In Figure 1d for chromosome IV, a particularly pseudogene-rich region is notable towards the start of the chromosome. Variation in the proportion of pseudogenes between chromosomes may be due to specific gene families, or perhaps recently defunct families of genes. To investigate this local variation further, specific clusters of pseudogenes along the chromosomes were identified (Figure 2; see Legend for definition of clusters). The largest of these clusters has twenty-one members and occurs on chromosome IV. We examined the rank #1 cluster in detail (Figure 2 legend) and found that all of the pseudogenes in this cluster correspond to genes from uncharacterized families.

We looked for recurrent pairs of predicted pseudogenes along the chromosomes that might give a clue as to their general chromosomal distribution. The most frequent pair patterns are tabulated (Table 2). The highest-scoring pair is of the ubiquitin C-terminal hydrolase family (representative F07A11.4) and an uncharacterized family (F11D11.10). The large clusters of predicted pseudogenes (Table 2) may in part arise from remnants of dead transposable elements that have not yet been documented; however, none of these pairs are indicative of this. Seven of the top ten pairs comprise a gene from a characterized family alongside an uncharacterized one; the other three are duplicated pairs of proteins from uncharacterized families.

Disablements and composition of ΨG

The obvious disablements for ΨG (frameshift or premature stop codon) are tallied in Table 4. A high proportion of ΨG has only one disablement over the length of genomic sequence aligned (48%). This may indicate an evolutionarily young pseudogene population that is rapidly deleted from genomic DNA. In general, non-coding frameshifts (of either one or two bases) and premature stop codons are approximately evenly represented in the pseudogene fragments detected (Table 4).

We measured the amino-acid composition of the Wormpep18 protein complement and the implied amino-acid composition of both ΨG and random genomic sequence (Figure 3). On a residue-to-residue basis, the amino-acid composition of ΨG is intermediate between the composition of random genomic sequence and the composition of the Wormpep18 proteins (Figure 3), being closer to random than to Wormpep18 (14 out of 20 residues). The residues that differ most in composition between ΨG and Wormpep18 are Phe, Ala, Asp and Glu. The increase in composition of Phe for ΨG relative to Wormpep18 is particularly interesting as the number of codons for this residue is small (two, TTT and TTC) (Figure 3). The fact that the composition values for Phe and Lys are elevated in the ΨG and random compositions relative to Wormpep18 is perhaps related to an underlying trend for local A/T mononucleotide repeats in the genome (data not shown). Also, Lys is preferred to the physico-chemically similar Arg in the *C. elegans* proteome even though the former has only two codons, compared with six for the latter (Figure 3 and Ref. [27]).

Distribution in terms of gene paralog families

We clustered the genes in the ribbon worm genome into gene paralog families (an example of a paralog family is illustrated in Figure 4). These paralog families are named for their particular paralog representatives (Table 3). We examined the prevalence of genes and pseudogenes for these families. Clearly, the estimated number of pseudogenes per family is not correlated with the number of genes per family (Table 3). Five of the top ten paralog gene families when ranked by number of pseudogenes are uncharacterized; only one of the gene families is in common with the top ten for gene families when ranked by the total number of genes (B0334.7, a 7TM receptor family). The occurrence of the reverse-transcriptase family in the list for pseudogenes may be due to the occurrence of an unknown transposable element. The rankings are similar for the EST-matched genes (ΨG_E) (Table 3a).

Figure 5 shows the number of genes in a family plotted versus the number of pseudogenes related to this family. One can clearly see that for large numbers of genes and pseudogenes there are many outliers from the overall ratio. There are some notable examples of nearly extinct families of genes. For example, the gene D1022.6 is a seven-transmembrane receptor that is homologous to opsins in other organisms, and only distantly similar to another gene in the ribbon worm (F57H12.6). However, there is an extensive array of pseudogenes for this gene (Table 3a), perhaps indicating a disused line of chemoreceptors [20, 21].

In addition, we found 150 pseudogenic homology fragments to representative sequences from the PROTOMAP database that were not detected to have homology to a worm protein (Table 3b). These either result from horizontal transfer or have diverged too far for the homology to their parent worm protein to be detected, or perhaps they are even remnants of gene families that have completely died out in the ribbon worm. The top match is a hypothetical protein from yeast (yja7_yeast), which has no other reported homologs (Table 3b).

Protein 'pseudofolds'

The proteins encoded by the ribbon worm genome have previously been assigned to globular protein domain folds from the SCOP database and assessed for the presence of transmembrane segments [23, 28]. Where possible, we assigned one of the known protein folds to each identified pseudogene based on standard approaches. In particular, for every pseudogene, the structural assignments of its closest gene homolog were considered as implied structural assignments (see *Methods*). Then we ranked the pseudogenes in terms of these implied structural assignments or 'pseudofolds' (Figure 6).

The prevalence of different globular folds is clearly different for the gene and pseudogene populations, although four folds occur in both top-ten lists (Figure 6). The immunoglobulin-like fold, which is in the all–β folding class, is the most prevalent in *G*, yet is only the eleventh ranking fold for *ΨG*. This fold is much more abundant in the worm than in any other completely sequenced organism [23]. The most common pseudofold for *ΨG* is the phosphoglycerate-mutase-like fold, which is an α/β class fold. This fold is associated with enzymatic proteins involved in carbohydrate metabolism. It is equally prevalent in all the major domains of life, though it is absent in certain select bacterial lineages [29]. Overall, there is a moderate decrease in assignability to a SCOP domain for the pseudogene population (16% have at least one assignment) compared to the gene population (24%). This may be due to truncation or deletion of genomic DNA.

We also extrapolated the transmembrane protein predictions [23] of the closest gene homolog to each pseudogene. The proportion of pseudogenes corresponding to a predicted transmembrane protein is about the same in *ΨG* (20%) as in *G* (22%). The proportion of seven-transmembrane proteins is small in both populations (G 2.3%, *Ψ*G 1.5%). In addition, outside of the Wormpep18 pseudogenes and pseudogene fragments, transmembrane helices were assigned on six-frame translations of the raw genomic sequence to locate other regions that are transmembrane-protein-like and pseudogenic (see *Methods* for details). There is a small number of such pseudogenic transmembrane segments with 4 or more predicted transmembrane helices (174 in total), that may be deceased transmembrane protein genes.

Conclusions

Our goal in this study was to provide an initial estimate of the size, distribution and characteristics of the pseudogene population for the genome of a eukaryote, *Caenorhabditis elegans*. We have found 3,814 homology fragments to confirmed and predicted genes in the worm genome (about 1 for every 5 genes) that appear to be pseudogenic. About a quarter of these (997) are for EST-matched genes. This figure may be an over-estimate due to inclusion of dead copies of transposable elements, or of under-predicted genes with disablements that are due to sequencing errors. Contrarily, it may be an under-estimate due to disregard for pseudogenes with only the less likely disablements, such as a damaged splicing signal.

We found few pseudogenes that are apparently due to processing in the ribbon worm genome. This is in marked contrast to the situation for the human genome, where 80% of the pseudogenes are thought to be processed [26].

The distribution of the proportion of pseudogenes relative to genes for different gene families is notably uneven, indicative of a highly dynamic genome. There are some striking examples of almost 'dead' gene families such as for the chemoreceptor (D1022.6), which is only distantly related to other chemoreceptors in the genome. Such genes or gene families may have recently fallen out of usage due to removal of the evolutionary pressure for their conservation, or due to recent functional redundancy with another gene family. This may partly explain why fewer pseudogenes occur for genes / gene families that are EST-matched. There are also some notable disparities in the relative rankings for numbers of folds and 'pseudofolds': the top ranking pseudofold for ΨG (the phosphoglycerate mutase fold) is much less abundant in the protein population.

There are more pseudogenes relative to genes on the arms of the chromosomes. This suggests that more duplications at the ends of the chromosomes tend to produce unusable genes. This may be because the arms of chromosomes undergo more recombination relative to the overall rate of genomic DNA loss. These areas may be thus more 'unreliable' for encoding genes and functions, but conversely are more likely to spawn new proteins. This may also explain the sparser occurrence of genes homologous to other organisms on the arms of the chromosomes [25]. There is general agreement in the chromosomal distribution for pseudogenes between the complete data set and the subset of EST-matched genes.

Acknowledgements

Thanks to Dr. Nathan Bowen (University of Atlanta) for LTR retrotransposon data in the ribbon worm genome and to Richard Durbin (Sanger Centre, UK) for helpful advice.

Methods

Digging in the ribbon worm genome for pseudogenes

We downloaded the following data from the Sanger sequencing center ftp site (ftp://www.sanger.ac.uk): the complete sequences of the six ribbon worm chromosomes, the most current Wormpep18 protein sequence database and GFF data files with annotations for genes and other genomic features. The *C. elegans* genome sequence data is constantly updated and certain regions will undoubtedly be revised in future versions; it should be stressed therefore that our survey results here are just an initial estimate of the extent of a pseudogene population. After Wormpep18 was initially masked for low complexity regions with the program SEG [30], the sequence alignment programs TFASTX and TFASTY [31] were used to compare the complete Wormpep18 against the worm genome. A list of representatives for sequence clusters from PROTOMAP [32] was also compared against the 99-megabase worm genomic DNA. Initial significant matches of the protein sequences to the genomic DNA (with e-value <= 0.01) were filtered for overlap with other annotations such as exons of genes, tandem repeats and transposable elements. We masked for transposable elements by comparing a library of sequences for reported (retro)transposons against the complete *C. elegans* genome sequence (including the Tc DNA transposons, the Rte-1 retrotransposon and LTR retrotransposons [33-35]). Sequences were checked for an obvious (sequence-length dependent) coding disablement (*i.e.*, either a frameshift or a premature stop codon) indicative of a pseudogene.

Prevention of over-counting for adjacent matches

Some of these initial matches may correspond to the same pseudogene. Therefore, to avoid over-counting for these worm protein matches, the initial matches were further aligned. The genomic DNA fragment *f* corresponding to each matching protein *p* was extracted. The predicted genomic sequence *g* for each paralog of the initial matching worm protein in the Wormpep18 database was aligned against *f*. The length of top-matching genomic sequence (g_{top}) relative to the fragment *f* gives an interval on the genomic DNA within which other less significant matches *f* can be discarded. This second alignment stage insures that two or more initial consecutive matches of a

Wormpep18 protein to genomic DNA are not counted as separate pseudogenes. The gene for g_{top} was also used as the final assignment as the closest homolog/paralog for a particular pseudogene.

The final list of pseudogenes was augmented with a small list of 332 pseudogenes annotated in the Sanger Center data to give the total pseudogene population (denoted ΨG).

Processed pseudogenes

We developed a heuristic to assess whether a pseudogene was processed. We estimated whether a pseudogene was processed by looking for 'exon seams' in the DNA segment *f*. An exon seam is a short stretch of coding sequence that would not be found uninterrupted in the genomic DNA without processing. We found that ten amino acids was a suitable length for an exon seam. If all but one of the exon seams for *g* of any paralogous protein *p* are found in the translation of *f* then the pseudogene is identified as a *possible* processed pseudogene. Processed pseudogenes have a polyadenine tract immediately 3' to their protein homology segment [2]. Polyadenosine tracts are added during messenger RNA processing and are usually between 50 and 200 nucleotides long. Therefore, in addition, we analysed a 50-nucleotide stretch immediately 3' to the pseudogene fragments found in the genomic DNA for any evidence of an elevated adenine content relative to the overall distribution of polyadenine content for predicted genes in the same region.

Clustering of Wormpep18 proteins

The 18,576 proteins on Wormpep18 were clustered using a modification of the algorithm of Hobohm *et al.* [36] for deriving representative lists of protein chains. Pairwise alignment using the FASTA algorithm [31] was performed to compare proteins. Two proteins were judged similar if they had an e-value for alignment <= 0.01. Clusters are formed in increasing order of the number of relatives that a sequence has in order to minimize false linkage. These clusters are termed *paralog families*. Each cluster is named after its representative Wormpep18 protein. Genes with no relatives according to this method are termed *singleton genes*.

Fold assignments

For the worm proteome, matches to SCOP domains and to transmembrane proteins are extrapolated onto Wormpep18 from assignments made previously on Wormpep17 proteins [23]. For the pseudogene complement, implied assignments to SCOP domains and transmembrane proteins are taken from the closest matching Wormpep18 protein for each individual pseudogene or pseudogene fragment.

In addition, we performed transmembrane helix prediction directly on six-frame translations of the raw genomic DNA using a hydropathy scale and 20-residue window as described in previous work [23, 29]. Based on an analysis of the distribution of length of interhelical segments in existing membrane protein structures, we joined two predicted transmembrane helices into the same 'exon' if they were separated by less than 40 amino acids. We only flagged the resulting assemblage as a pseudogene if it contained a single stop codon in one of the predicted transmembrane helices. These predicted transmembrane protein regions are masked for overlap with other described genomic features as for the pseudogene homology matching.

EST-confirmed subset of worm gene sequences

Some of the gene sequences in the Sanger Centre worm genome data are noted as matched to ESTs or full-length cDNA. A further set of EST- and cDNA-confirmed worm gene structures is available from the Intronerator database (http://www.cse.ucsc.edu/~kent/intronerator; [37]). We merged these two sets of notations and derived two sets of EST-verified genes/proteins. Firstly, the set of genes with at least one verifying EST were compiled (G_E). Secondly, G_E was expanded by including all of the paralogs of G_E proteins (G_P).

Data on website

We have constructed a web site http://bioinfo.mbb.yale.edu/genome/worm/intergenic for browsing the pseudogene annotations, along with other genomic features downloaded from the Sanger Centre website. The data can be viewed either by searching for a particular ORF or protein name, by viewing the region around an ORF, or simply by viewing a specified range in the chromosome. The sense and alignment score of all pseudogenes is displayed, and the genomic sequences of aligned segments (along with their amino acid translations) are viewable. We have also linked the results to a variety of available internal and external resources including online databases and structural annotations.

References

1. Weiner, A.M., P.L. Deininger, and A. Efstratiadis. (1986) *Annu. Rev. Biochem.*, **55**, 631-661.
2. Vanin, E.F. (1985) *Ann. Rev. Genet.*, **19**, 253-272.
3. Mighell, A.J., N.R. Smith, P.A. Robinson, and A.F. Markham. (2000) *FEBS Letts.*, **468**, 109-114.
4. Korneev, S.A., J.-H. Park, and M. O'Shea. (1999) *J. Neurosci.*, **19**, 7711-7720.
5. Olsen, M.A. and L.E. Schechter. (1999) *Gene*, **227**, 63-69.

6. Feenstra, M., J. Bakema, M. Verdaasdonk, E. Rosemuller, J. van den Tweel, P. Slootweg, R. Weger, and M. Tilanus. (2000) *Gen. Chrom. Cancer*, **27**, 26-34.
7. Fujii, G.H., A.M. Morimoto, A.E. Berson, and J.B. Bolen. (1999) *Oncogene*, **18**, 1765-1769.
8. Currie, P.D. and D.T. Sullivan. (1994) *Genetics*, **138**, 353-363.
9. Sullivan, D.T., W.T. Starmer, S.W. Curtiss, M. Menotti, and J. Yum. (1994) *Mol. Biol. Evol.*, **11**, 443-458.
10. Petrov, D., E. Lzovzkaya, and D. Hartl. (1996) *Nature*, **384**, 346-349.
11. Gojobori, T., W.H. Li, and D. Graur. (1982) *J. Mol. Evol.*, **18**, 360-369.
12. Li, W.H., C.I. Wu, and C.C. Luo. (1984) *J. Mol. Evol.*, **21**, 58-71.
13. Averof, M., A. Rokas, K.H. Wolfe, and P.M. Sharp. (1999) *Science*, **287**, 1283-1285.
14. Gu, X. and W.-H. Li. (1995) *J. Mol. Evol.*, **40**, 464-473.
15. Ophir, R. and D. Graur. (1997) *Gene*, **205**, 191-202.
16. Petrov, D., T.A. Sangster, J.S. Johnston, D.L. Hartl, and K.L. Shaw. (2000) *Science*, **287**, 1060-1062.
17. Ota, T. and M. Nei. (1995) *Mol. Biol. Evol.*, **12**, 94-102.
18. Goncalves, I., L. Duret, and D. Mouchiroud. (2000) *Genome Res.*, **10**, 672-678.
19. The C. elegans Genome Sequencing Consortium (1998) *Science*, **282**, 2012-2018.
20. Robertson, H.M. (1998) *Genome Res.*, **8**, 449-463.
21. Robertson, H.M. (2000) *Genome Res.*, **10**, 192-203.
22. Gerstein, M. (1997) *J. Mol. Biol.*, **274**, 562-576.
23. Gerstein, M., J. Lin, and H. Hegyi. (2000) *Pac. Symp. Biocomputing*, **5**, 30-42.
24. Jansen, R. and M. Gerstein. (2000) *Nucleic Acids Res.*, **28**, 1481-1488.

25. Chervitz, S.A., et al. (1998) *Science*, **282**, 2016-2022.

26. Dunham, I., et al. (1999) *Nature*, **402**, 489-495.

27. Nishizawa, M. and K. Nishizawa. (1998) *J. Mol. Evol.*, **47**, 385-393.

28. Murzin, A.G., S.E. Brenner, T. Hubbard, and C. Chothia. (1995) *J. Mol. Biol.*, **247**, 536-540.

29. Gerstein, M. (1998) *Proteins*, **33**, 518-534.

30. Wootton, J.C. and S. Federhen. (1996) *Methods Enzymol.*, **266**, 554-571.

31. Pearson, W.R., T. Wood, Z. Zhang, and W. Miller. (1997) *Genomics*, **46**, 24-36.

32. Yona, G., N. Linial, and M. Linial. (2000) *Nucleic Acids Res.*, **28**, 49-55.

33. Youngman, S., H.G.A.M. van Luenen, and R.H.A. Plasterk. (1996) *FEBS Letters*, **380**, 1-7.

34. Bigot, Y., C. Auge-Gouillou, and G. Periquet. (1996) *Gene*, **174**, 265-271.

35. Bowen, N. and J. McDonald. (1999) *Genome Res.*, **9**, 924-935.

36. Hobohm, U., M. Scharf, R. Schneider, and C. Sander. (1992) *Protein Sci.*, **1**, 409-417.

37. Kent, W.J. and A.M. Zahler. (2000) *Nucleic Acids Res.*, **28**, 91-93.

38. Altschul, S.F., T.L. Madden, A.A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D.J. Lipman (1997) *Nucleic Acids Res.*, **25**, 3389-3402.

Figure Legends

Figure 1: The estimated chromosomal distribution of pseudogenes. Each panel depicts the distribution of genes (left hand side) and pseudogenes (right hand side) for the chromosomes I, II, III, IV, V, X. The EST-matched subsets for each chromosome are binned as a dark grey bar with the remainder of the genes pseudogenes as a light grey bar. The bin size is 250,000 bases. The axis for number of pseudogenes is scaled by two (*X2*) relative to the same axis for genes. The total

estimated sizes of the chromosomal populations of pseudogenes are as follows (the columns are chromosome name, total number of genes, total number of exons for genes, total number of pseudogenes and the proportion of 'dead' gene copies) :-

Chromosome	$\mathbf{|G_{chromosome}|}$	**|Exons|**	$\mathbf{|\Psi G_{chromosome}|}$	$\mathbf{|\Psi G_{chromosome}|}$ / $\mathbf{[|G_{chromosome}| + |\Psi G_{chromosome}|\,]}$
I	2645	17641	557	0.17
II	3338	19931	536	0.14
III	2347	15243	445	0.16
IV	2757	16824	811	0.23
V	4737	26756	953	0.17
X	2684	19508	512	0.16

Figure 2: The largest clusters of pseudogenes are on chromosome IV. A pseudogene cluster is defined as a contiguous group of pseudogenes along a chromosome. The clusters are 'smoothed' so that genes that are *immediately* adjacent to any pseudogene are ignored. The approximate positions of each of the largest six pseudogene clusters in the ribbon worm genome are marked by a bar (a longer bar indicates the largest cluster) with the following information: the rank of the cluster in terms of size (largest is rank #1), total number of pseudogenes, the range of the pseudogene cluster along the chromosome (in chromosomal coordinates). The chromosome is drawn with coordinates to scale. The members of cluster #1 are:

W06A11.3 (4 times), T05A7.9 (4), Y71F9B.E (2), ZK402.5 (1), Y54B9A.A (1), Y48E1B.11 (1), Y39A3CL.B (1), Y32B12B.1 (1), Y17G7B.10B (1), F49E12.2 (1), F38H4.1 (1), F22B3.7 (1)

Figure 3: Composition for the ΨG data set. The amino-acid composition of the Wormpep18 database is compared to the implied amino-acid composition of random genomic sequence and the ΨG population. The percentage composition for each of the twenty amino acids is graphed in

decreasing order of the implied amino acid composition in the pseudogene set. In the bottom part of the figure, the *spread value* for each amino acid composition is indicated by a bar. This is defined as ($|w\text{-}p|$ + $|p\text{-}r|$) / p , where w is the amino-acid composition value for the Wormpep18 proteins, r is the implied composition for random genomic sequence and p is the implied pseudogene composition. The asterisk (*) in this graph represents the termination codons. The number of codons for each amino-acid type is written below the one-letter code for the residue.

Figure 4: An example of a paralog family with associated pseudogenes. The positions of genes for the paralog family whose representative is the sequence C02F4.2, are indicated by grey ovals (totalling 40). The pseudogenes are marked with black ovals (totalling 4). A pseudogene fragment (ΨC02F4.2) from chromosome II is shown along with an example of a gene from this paralog family W09C3.6 (which is for a serine/threonine protein phosphatase PP1) with the homologous segment underlined. The pseudogene is interrupted by a frameshift relative to this gene (marked by a # symbol). The corresponding sequence in the gene paralog is boxed in black. This corresponds to one exon of the gene paralog. The stop codon of the gene is marked by an asterisk (*).

Figure 5: Plot of the number of genes in a paralog family (G_{family}) *versus* the number of pseudogenes in a paralog family (ΨG_{family}). The families from the G_E set are marked as grey filled points, with the remainder as unfilled points. The lines indicate the overall ratio of the number of genes to the number of pseudogenes for the whole genome and for the G_E subset. Large families that are outliers from this overall trend are labeled with the name of their family representative.

Figure 6: The folds and pseudofolds in the ribbon worm genome. The SCOP domain matches (part *(a)* of the figure) are extrapolated onto Wormpep18 from assignments made previously on Wormpep17 proteins [23]. 'Pseudofold' assignments (part *(b)*) are taken from the closest matching gene paralog for each pseudogene. The columns are as follows: Rank for folds or pseudofolds (with total numbers in brackets); corresponding rank for pseudofolds or folds; a fold cartoon; the representative domain, the SCOP 1.39 domain number and a brief description of the fold. The fold cartoons are coloured in a sliding gradient from blue for the N-terminus to red for the C-terminus.

Table 1: Overall statistics for ΨG

	Category	**Total number**	**Number for genes with EST match**	**Genes with EST match as percentage of *Category***	**Number for genes in paralog families with EST match**	**Genes in paralog families with EST match as percentage of *Category***
Genes	Total	18,576 (G)	7,829 (G_E)	42%	13,417 (G_P)	72%
	Singletons	5,913	2,788	47%	---	---
Pseudogenes and pseudogene fragments	Total	3,814 (ΨG)	997 ($ΨG_E$)	26%	2,729 ($ΨG_P$)	72%
	Singletons	637 (17% of ΨG)	233	36%	---	---
	Intronic pseudogenes *	1,155 (30% of ΨG)	351	30%	704	61%

* The numbers of sense and antisense intronic pseudogenes are 564 (49%) and 591 respectively.

Table 2: Most common pair patterns for predicted pseudogenes along chromosomes *

Pair of pseudogenes**	**Number of occurrences**
ΨF07A11.4, ΨF11D11.10	46
ΨC05C8.8, ΨC50C3.8	46
ΨF11D11.10, ΨF07A11.4	40
ΨAC3.2, ΨC50E10.2	30
ΨC50C3.8, ΨC05C8.8	26
ΨAC3.1, ΨAC3.1	24
ΨC05C8.8, ΨC05C8.8	22
ΨB0334.7, ΨB0334.7	20
ΨY48E1B.11, ΨC05C8.8	18
ΨF15D4.3, ΨC50C3.8	18

*Uncharacterized proteins are underlined.

**Each pseudogene or pseudogene fragment is named according to its closest gene paralog, with the prefix *Ψ*.

Table 3:

(A) Top paralog families for ΨG and G *

Rankings for ΨG			**Rankings for G**		
Name of family representative	**ΨG_{family}**	**Note**	**Name of family representative**	**G_{family}**	**Note**
C50C3.8 ^{E}	90	uncharacterized	**B0280.8** ^{E}	216	Ligand-binding domain of Nuclear Hormone receptor
C05C8.8 ^{E}	90	uncharacterised	**B0334.7** ^{E}	193	7TM receptor
D1022.6 ^{E}	85	7TM receptor	**B0213.7** ^{E}	188	7TM receptor
B0334.7 ^{E}	75	7TM receptor	**B0205.7** ^{E}	124	Casein-kinase protein kinase
F11D11.10	67	uncharacterised	**B0047.1** ^{E}	93	MATH domain
F36H9.1	64	uncharacterised	C03A7.3	70	7 TM receptor
Y59A8B.O ^{E}	58	SnRNP-associated splicing factor	**AH6.1** ^{E}	70	Guanylyl cyclase / receptor tyrosine kinase
B0281.2 ^{E}	58	Reverse transcriptase	**B0213.10** ^{E}	70	Cytochrome P450
F07A11.4 ^{E}	57	An ubiquitin C-terminal hydrolase domain family	**B0207.1** ^{E}	70	Protein tyrosine phosphatase
M151.1 ^{E}	48	uncharacterised	**AC3.2** ^{E}	68	UDP-glucosyltransferase

*Paralog families for EST-matched proteins are in bold and are labeled with a superscript E. B0334.7 is underlined as it is common to both lists.

(B) Other pseudogenic homology fragments that match a PROTOMAP family representative but with no detected homology to a WormPep protein

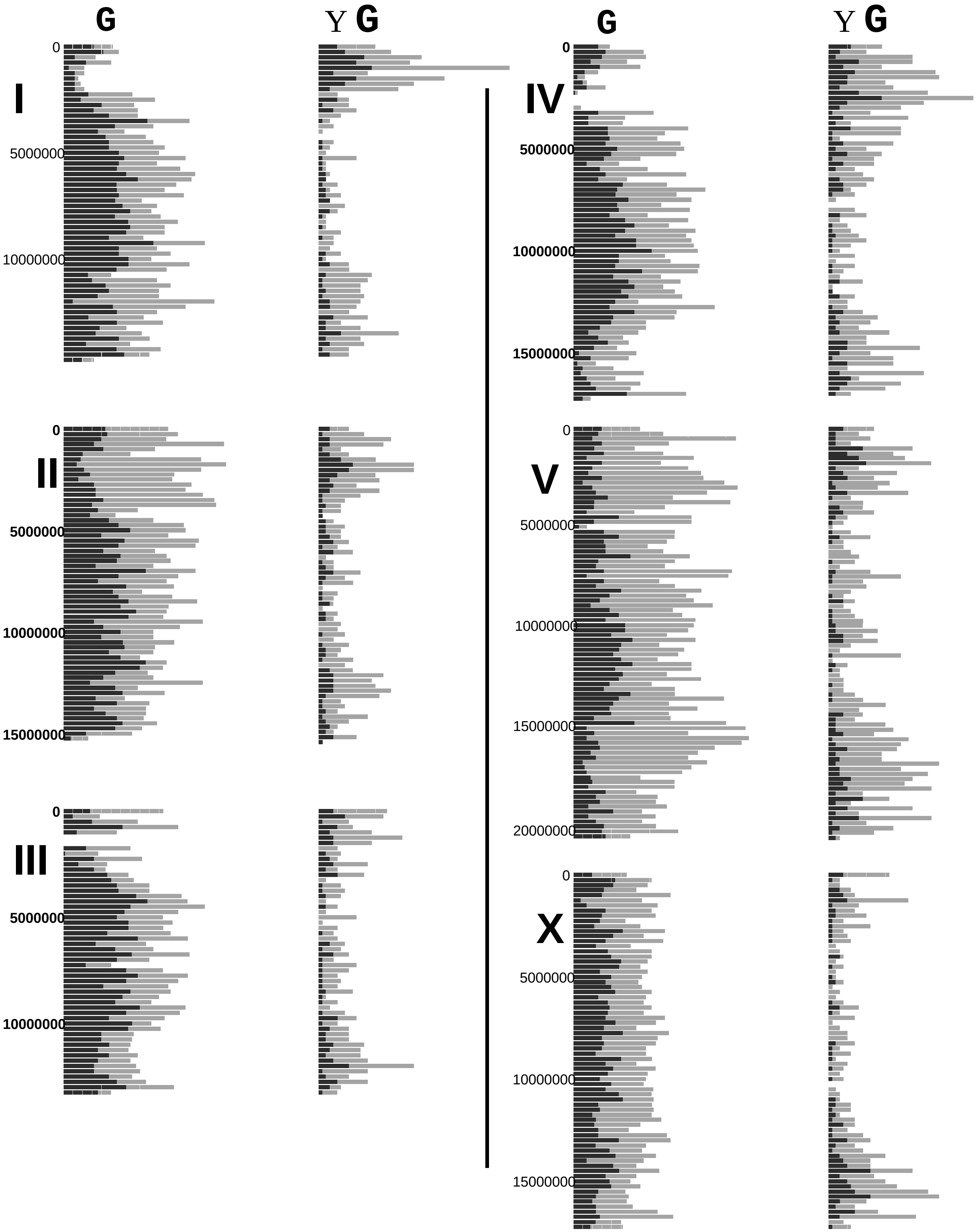
Rank	Name of PROTOMAP family representative	Number of matches	Organism of closest match*	Note on family representative
#1	YJA7_YEAST	7 *******	Yeast	Hypothetical protein in yeast
#2 =	XPD_MOUSE	5 *****	Human	Xeroderma pigmentosum group D complementing protein
#2 =	CPSA_BOVIN	5 *****	Bovine	Cleavage and polyadenylation specificity factor
#4 =	THB_RANCA	4 ****	Xenopus laevis	Thyroid hormone receptor beta
#4 =	SEX_HUMAN	4 ****	Human	SEX gene
#4 =	MDR1_RAT	4 ****	Drosophila	Multidrug resistance protein 1
#7 =	YVFB_VACCC	3 ***	Vaccinia virus	Hypothetical vaccinia virus protein
#7 =	VHRP_VACCC	3 ***	Drosophila	Host range protein from vaccinia
#7 =	IF4V_TOBAC	3 ***	Human	Eukaryotic initiation factor 4A
#7 =	ACRR_ECOLI	3 ***	E. coli	Acrab operon repressor

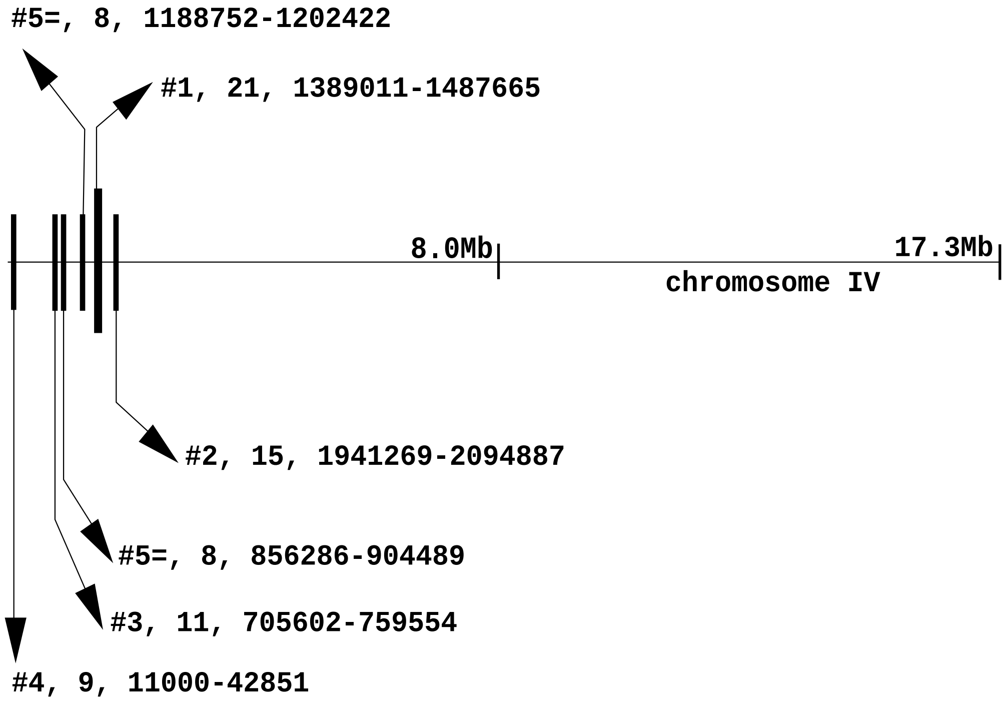
*Determined by a database search with the PSI-BLAST alignment program [38].

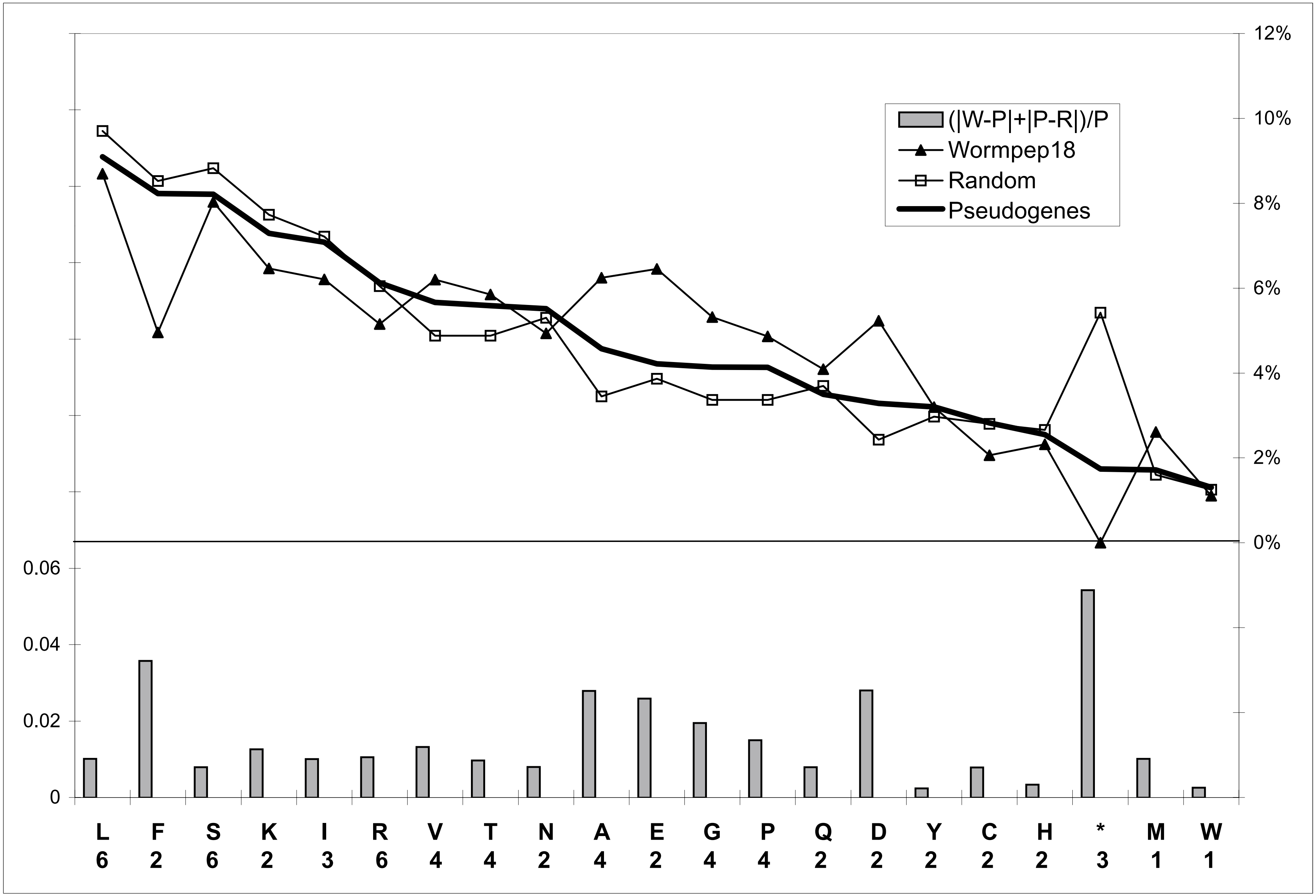
Table 4: Statistics for obvious disablements in pseudogenes and pseudogene fragments

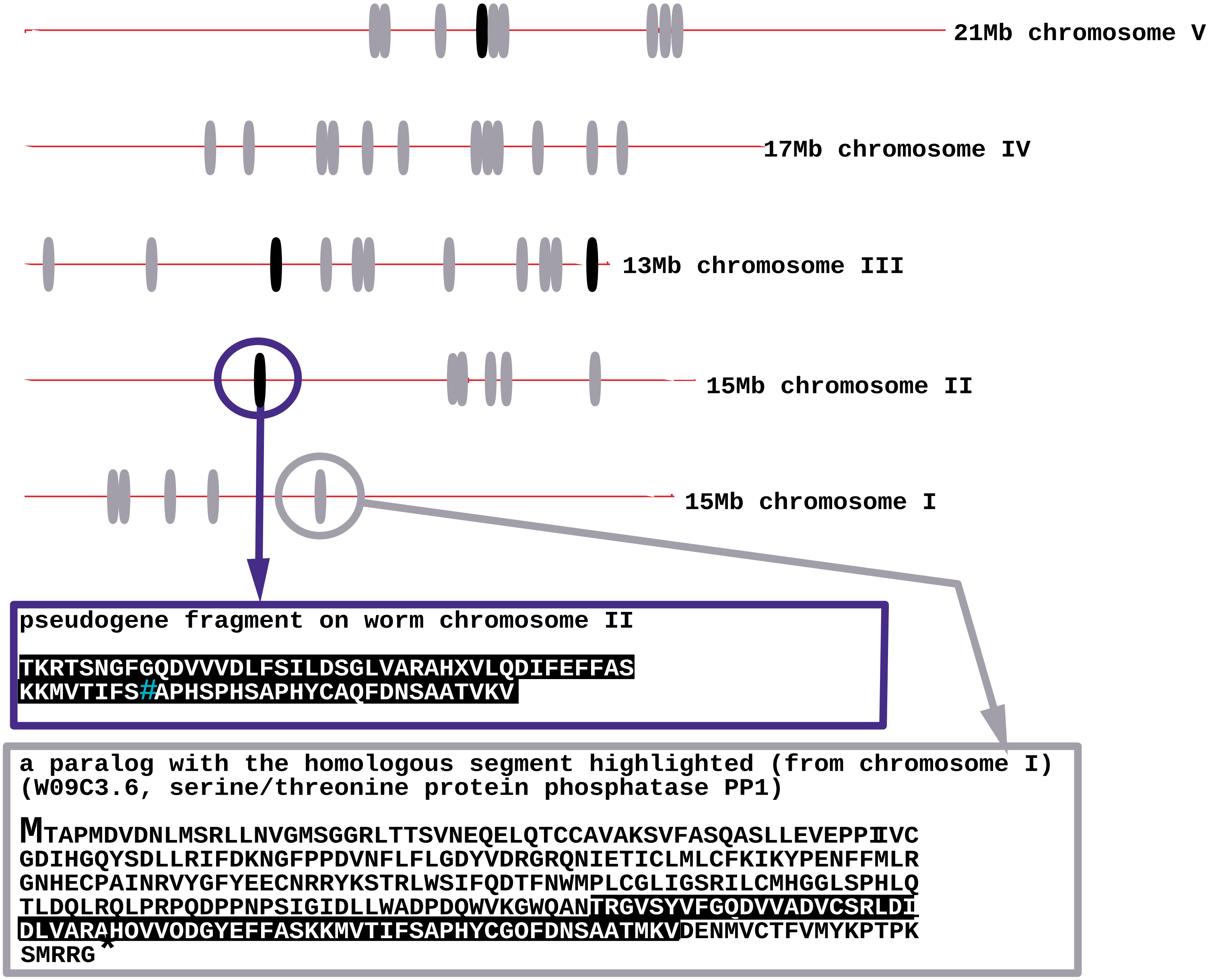
Category*	**Numbers**	**Total**
Frameshifts	2,001 (mononucleotide) 1,459 (dinucleotide)	3,460
Premature stop codons	---	4,049
Sequences with one disablement	872 (frameshift) 797 (premature stop codon)	1,669

* This data is only for the ΨG population without the additional Sanger Centre pseudogene data.


G
Y G
G
Y G
I
0
5000000
10000000
II
0
5000000
10000000
15000000
III
0
5000000
10000000
IV
0
5000000
10000000
15000000
V
0
5000000
10000000
15000000
20000000
X
0
5000000
10000000
15000000


#5=, 8, 1188752-1202422
#1, 21, 1389011-1487665
8.0Mb
17.3Mb
chromosome IV
#2, 15, 1941269-2094887
#5=, 8, 856286-904489
#3, 11, 705602-759554
#4, 9, 11000-42851


(|W-P|+|P-R|)/P
Wormpep18
Random
Pseudogenes
12%
10%
8%
6%
4%
2%
0%
0.06
0.04
0.02
0
L 6
F 2
S 6
K 2
I 3
R 6
V 4
T 4
N 2
A 4
E 2
G 4
P 4
Q 2
D 2
Y 2
C 2
H 2
* 3
M 1
W 1

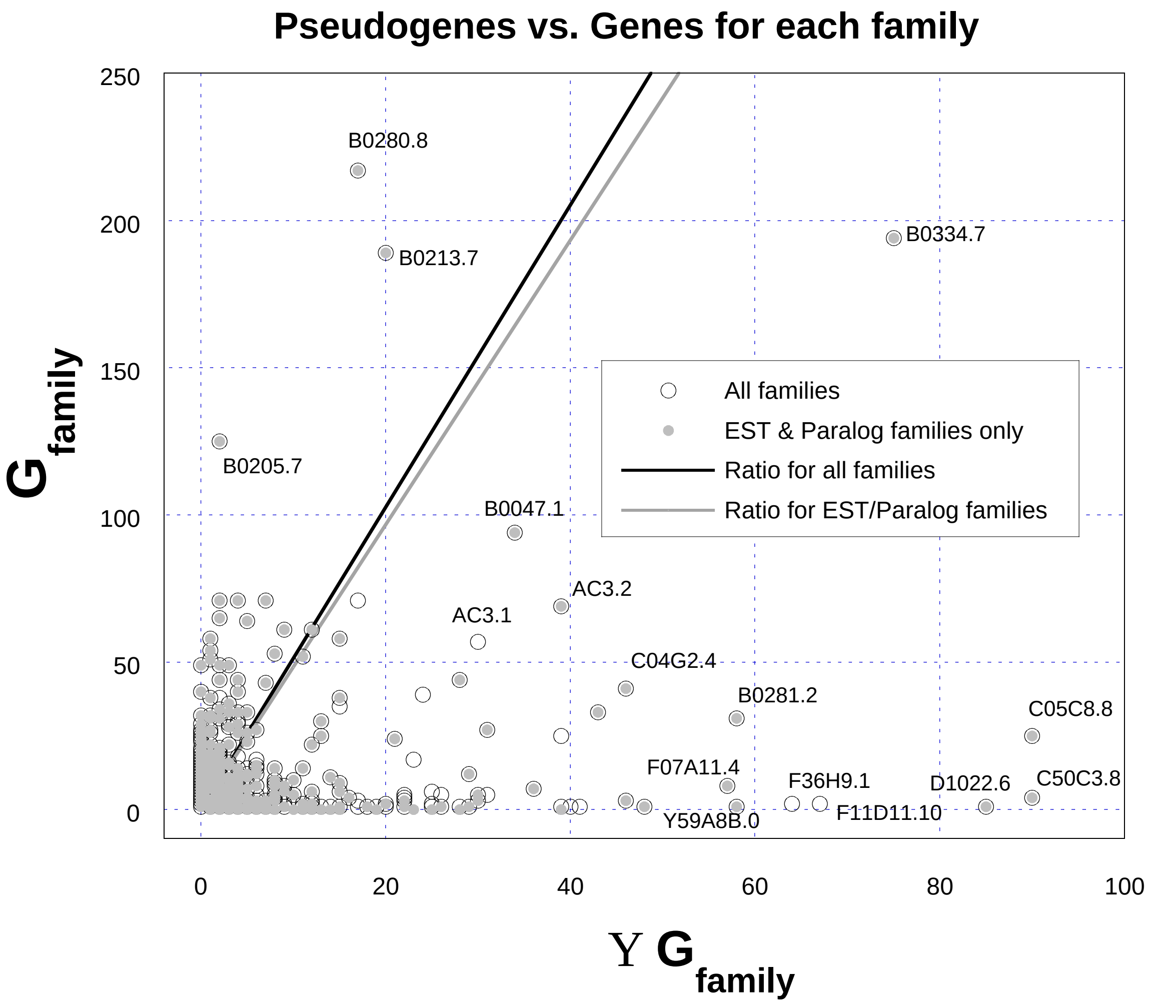


pseudogene fragment on worm chromosome II

TKRTSNGFGQDVVVDLFSILDSGLVARAHXVLQDIFEFFAS
KKMVTIFS#APHSPHSAPHYCAQFDNSAATVKV

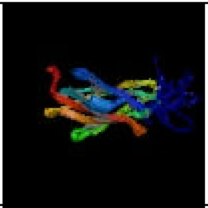
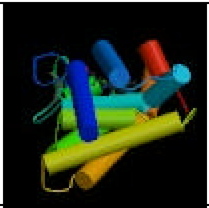
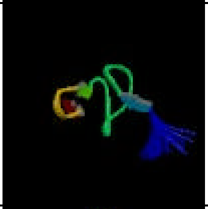
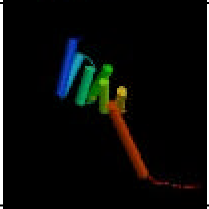
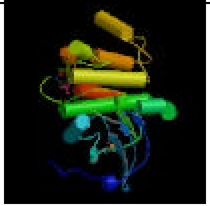
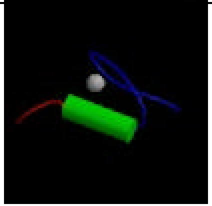
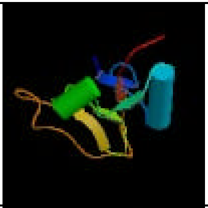
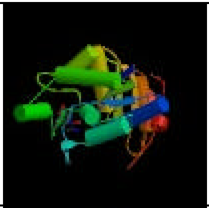
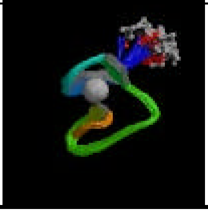
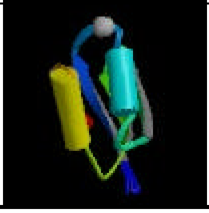
a paralog with the homologous segment highlighted (from chromosome I)
(W09C3.6, serine/threonine protein phosphatase PP1)

MTAPMDVDNLMSRLLNVGMSGGRLTTSVNEQELQTCCAVAKSVFASQASLLEVEPPIIVC
GDIHGQYSDLLRIFDKNGFPPDVNFLFLGDYVDRGRQNIETICLMLCFKIKYPENFFMLR
GNHECPAINRVYGFYEECNRRYKSTRLWSIFQDTFNWMPLCGLIGSRILCMHGGLSPHLQ
TLDQLRQLPRPQDPPNPSIGIDLLWADPDQWVKGWQANTRGVSYVFGQDVVADVCSRLDI
DLVARAHQVVQDGYEFFASKKMVTIFSAPHYCGQFDNSAATMKVDENMVCTFVMYKPTPK
SMRRG*


Pseudogenes vs. Genes for each family
250
200
150
100
50
0
G family
0
20
40
60
80
100
Y G family
B0280.8
B0213.7
B0334.7
B0205.7
B0047.1
AC3.2
AC3.1
C04G2.4
B0281.2
C05C8.8
F07A11.4
F36H9.1
D1022.6
C50C3.8
Y59A8B.0
F11D11.10
○ All families
● EST & Paralog families only
Ratio for all families
Ratio for EST/Paralog families

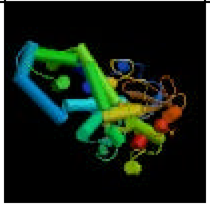
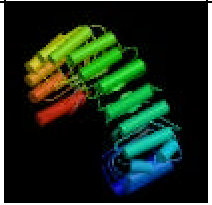
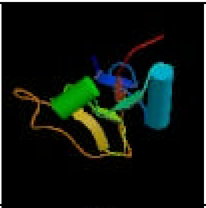
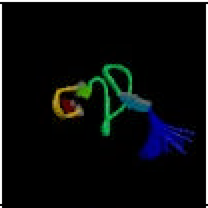
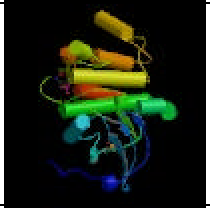
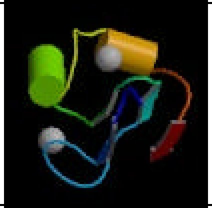
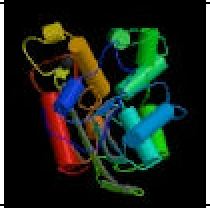
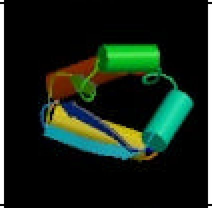
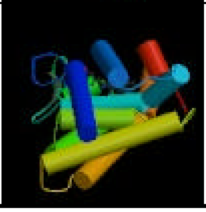
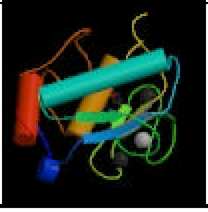
Part *(a)*

Fold Rankings for *G*

G Rank (Number of matches)	Y G Rank	Fold	Representative Domain, SCOP 1.39 Number, Description	G Rank (Number of matches)	Y G Rank	Fold	Representative Domain, SCOP 1.39 Number, Description
1 (769)	11		d1ajw__ 2.1 Immunoglobulin	6 (246)	5		d2lbd__ 1.95 Nuclear receptor ligand-binding domain
2 (555)	7		d1dec__ 7.3 Knottin	7 (243)	21		d1a17__ 1.91 Alpha/alpha superhelix
3 (434)	3		d3lck__ 5.1 Protein kinase	8 (227)	27		d1sp2__ 7.31 Classic zinc finger
4 (302)	2		d1tsg__ 4.105 C-type lectin	9 (215)	20		d1dai__ 3.29 P-loop NTP hydrolase
5 (274)	13		d1zfo__ 7.33 Gluco-corticoid receptor DNA-binding domain	10 (197)	17		d2aw0__ 4.34 Ferredoxin

Part *(b)*

Pseudofold rankings for ΥG

ΥG Rank (Number of matches)	ΥG Rank	Fold	Representative Domain, SCOP 1.39 Number, Description	Υ G Rank (Number of matches)	G Rank	Fold	Representative Domain, SCOP 1.39 Number, Description
1 (68)	49		d1ihp__ 3.48 Phospho-glycerate mutase	6 (39)	19		d2bnh__ 3.7 Leucine-rich repeat, right-handed beta/alpha superhelix
2 (51)	4		d1tsg__ 4.105 C-type lectin	7 (32)	2		d1dec__ 7.3 Knottin
3 (43)	3		d3lck__ 5.1 Protein kinase	8 (30)	15		d1bor__ 7.37 RING finger domain
4 (40)	11		d1cvl__ 3.56 Alpha/beta-hydrolase	9 (28)	33		d1a68__ 4.24 POZ domain
5 (40)	6		d2lbd__ 1.95 Nuclear receptor ligand-binding domain	10 (27)	28		d1mmq__ 4.52 Zincin