Digging Deep for Ancient Relics: A Survey of Protein Motifs in the Intergenic Sequences of Four Eukaryotic Genomes

Running title: Survey of Protein Pseudomotifs in Intergenic Region

ZhaoLei Zhang, Paul Harrison and Mark Gerstein[*, †]

Department of Molecular Biophysics and Biochemistry and [†]Department of Computer Science, Yale University, 266 Whitney Avenue, New Haven, CT 06520-8114

[*]Corresponding author

Tel: (203) 432 6105

Fax: (360) 838 7861

Email: Mark.Gerstein@yale.edu

Summary

We have examined conserved protein motifs in the non-coding, intergenic regions ("pseudomotif patterns") and surveyed their occurrence in the fly, worm, yeast and human genomes (chromosomes 21 and 22 only). To identify these patterns, we masked out annotated genes, pseudogenes and repeat regions from the raw genomic sequence and then compared the remaining sequence, in six-frame translation, against 1319 patterns from the PROSITE database. For each pseudomotif pattern, the absolute number of occurrences is not very informative unless compared against a statistical expectation; consequently, we calculated the expected occurrence of each pattern using a Poisson model and verified this with simulations. Using a *p*-value cut-off of 0.01, we found 67 pseudomotif patterns over-represented in fly intergenic regions, 34 in worm, 21 in human and 6 in yeast. These include the Zinc finger, leucine zipper, nucleotide-binding motif and EGF domain. Many of the over-represented patterns were common to two or more organisms, but there were a few that were unique to specific ones. Furthermore, we found more over-represented patterns in the fly than in the worm, although fly has fewer pseudogenes. This puzzling observation can be explained by a higher deletion rate in the fly genome. We also surveyed under-represented patterns, finding 23 in the fly, 12 in worm, 18 in human and 2 in yeast. If intergenic sequences were truly random, we would expect an equal number of over and under-represented patterns. The fact that for each organism the number of over-represented patterns is greater than the number of under-represented ones implies that a fraction of the intergenic regions consist of ancient protein fragments that, due to accumulated disablements, have become unrecognizable to conventional techniques for gene and pseudogene identification. Moreover, we find that in aggregate the over-represented pseudomotif patterns occupy a substantial fraction of the intergenic regions.

Keywords: genome, intergenic, pseudogene, protein motif, DNA loss

Introduction

In the genomes of higher organisms, only a small portion encodes for protein sequences, and of the non protein-coding sequence only a small fraction is known to have regulatory functions. In the human genome, it is estimated that 97 % of the sequence are non-protein coding [1, 2]. It has been a great challenge for biologists to understand the origin and possible functions of these intergenic sequences in the genome. Comparative alignment of non-coding sequences between human and mouse [3, 4], *C. elegans* and *C. Briggsae* [5], *D. melanogaster* and *D. virilis* [6] found that around 20 ~ 30% of the non-coding sequences in eukaryotic genomes are functionally constrained among evolutionally-related species, suggesting that these regions have important regulatory roles. Despite these progress, understanding the origin and function of intergenic sequences remains a final frontier in the post-genomic era.

Pseudogenes are non-functional copies of functional genes in the genome; they are often referred to as "dead genes" [7, 8] in the sense that they have become disabled in the course of evolution as results of nucleotide substitutions, insertions or deletions. However these pseudogenes have not deviated too far from their original sequences, so they can still be recovered by homology matching and the observation of simple disablements in the sequence, such as premature stop codons or frame shifts. Large populations of pseudogenes have been discovered in the genomes of the worm [9], yeast [10] and human [11].

It is logical to argue that there could be more ancient genes or gene fragments in the genome that, after becoming disabled and freed from selection pressure, have accumulated so many substitutions, insertions and deletions that they can no longer be recognized by homology matching methods. As an extension to our on-going research on pseudogenes [9-12], we would like here to survey the pseudogenic protein fragments that can be recovered from intergenic regions. In particular, comparing the occurrence of "living" and "dead" protein parts will directly address a number of interesting evolutionary questions: is the occurrence of a protein family in dead

proteins related to its occurrence in the live proteome? Or are there certain protein families that were more common in the past and are dying out of modern eukaryotic organisms? Do these decaying families have any similarity to proteins in supposedly "more primitive" bacteria and archea?

Since these ancient gene relics or "pseudomotifs" used to be part of protein coding sequences, we decided to search for protein fragments in the amino acid sequences predicted from raw intergenic sequences. It is important to choose the appropriate searching queries. PROSITE [13, 14] is a database of biologically significant protein patterns and profiles; it is maintained and frequently updated, and every pattern in it represents a conserved protein motif [15]. For each protein pattern from PROSITE, we did a comprehensive search in the intergenic region of the yeast, fly and worm genomes and the human chromosome 21 and 22, counting the number of occurrences of "pseudomotifs". Because of the small and fragmentary nature of many of the patterns, it is quite conceivable that many of the matches we found could have arisen purely by chance. Consequently, for a specific individual match, it is difficult to determine whether it was a real protein motif. However, the focus of our study is to survey the pattern occurrences in a collective and statistical way rather than on an individual basis. We designed a probabilistic framework so that for each PROSITE pattern we could calculate the expected number of occurrences based on a Poisson model, and consequently determine the statistical significance of the observed number of occurrences. Using a *p*-value cut-off of 0.01, we were able to find those patterns that have more than the expected number of occurrences (i.e., those that are over-represented) in the intergenic regions. We want to emphasize that we deleted all the easy matches, i.e. the genes, pseudogenes and repeats from the genome, and we still found significantly over-represented protein patterns. This suggests that at least a fraction of the intergenic sequences in the genome could have arisen from ancient protein-coding genes.

We like to consider our study as "genomic paleontology" because, just like paleontologists digging for animal fossils, we are searching for molecular fossils, i.e. ancient protein fragments. When paleontologists excavate a fossil site, it is likely that more recent and more complete fossil samples, perhaps a whole skeleton, would be found in the upper layer of the depository; this is analogous to the discovery of pseudogenes, which are more recent in evolution and bear more homology to functional genes. When the paleontologists dig deeper under the ground, they tend to find more ancient and more fragmented fossils (e.g. only an individual limb or a tooth), like our findings of ancient protein fragments.

Results And Discussion

Overall Many PROSITE Patterns are Over-represented

We used a *p*-value cut-off of 0.01 to select patterns that have an "unusual" number of occurrences relative to expectation. We divided the 1319 PROSITE patterns into three groups: over-represented, under-represented, and similar-to-expectation. This last category was for patterns that have statistically non-significant differences in occurrence relative to expectation -- i.e. those that have *p*-values ≥ 0.01, regardless of occurrence. While we believe a *p*-value cut-off of 0.01 is stringent enough, we also analyzed our data with a stricter *p*-value cut-off of 0.001; our conclusions were unaffected.

As can be seen in Table 1, we picked up some very strongly over-represented pseudomotif patterns, which have *p*-values as small as 1.0e-300. As summarized in Table 2 (A), the fly has the most over-represented patterns in its intergenic regions, the worm and the human (chromosome 21 and 22 only) have fewer and yeast has the least. The same trend can be seen for under-represented patterns as well. It is not surprising that for the majority of the PROSITE patterns, their occurrences are similar to expectation. The pseudomotif patterns we are looking for are very faint signals of ancient proteins; thus, it is expected the background noise level would be

significantly high. Furthermore, we have carefully removed all sources of known "signals" -- i.e. genes, pseudogenes, and repeats.

It is interesting to estimate the fraction of the present intergenic region that once coded for ancient protein. If we multiply the length of each over-represented pseudomotif pattern by the difference between the observed and expected number of occurrences, we find that significantly over-represented patterns occupy ~3% of the total intergenic region (Table 4). On average a PROSITE motif is ~14 amino acids in size. This is 12% of the size (120) of an average protein domain [16], implying that the total coverage of the domains associated with the motifs is ~26 % of the intergenic region in all four genomes.

Note that this result does not include the amount of intergenic DNA in "obvious" pseudogenes. Furthermore, it could be affected by the fact that the PROSITE database is an incomplete set of features and is heavily biased by human input. Nevertheless, based on our results it is reasonable to assume that at a significant part of the intergenic region, otherwise known as “junk DNA”, once were protein-coding genes.

More Patterns are Over-represented than Under-represented

Overall, more patterns are over-represented than under-represented in all four organisms. This trend is strongest in the fly and worm. It is significant that more over- than under-represented patterns are found; this indicates in a broad fashion the existence of non-random or protein-like features in intergenic regions. If the intergenic sequences were truly random, then we would expect an equal number of over- and under-represented patterns.

One possible explanation for these over-representing patterns is that they are simply repetitive genomic DNA sequences that happen to translate to PROSITE protein patterns. To rule out this hypothesis, we selected some prominent over-represented patterns, such as Zinc finger, Leucine zipper, ATP_GTP binding site, and examined the translated amino acid sequence of the actual intergenic matches. The results clearly show no dominant amino acid sequence among the

matches; in fact most of the matches have unique amino acid sequences (data on website). Thus we are assured that enrichment of the patterns is not result of DNA repetition.

Specific Over- and Under-represented Patterns

Table 1 lists the number of occurrences in the fly and worm intergenic regions for selected PROSITE patterns. The over-represented patterns are placed at the top of the table and sorted by *p*-value in ascending order. (As described in the Methods, we did simulations to verify that these patterns were indeed over-represented.) Under-represented patterns are at the bottom of the table and sorted by *p*-value in descending order. The statistically non-significant patterns are placed in the middle and sorted first by occurrences and then by *p*-value. By sorting the pseudomotif patterns this way, we can give each one a rank, from 1 to 1319, indicating its relative "level of enrichment" compared with the other patterns. Only the top 20 and bottom 5 of the patterns are shown here. The rest of the data are on our website.

There are some well-known protein motifs among the over-represented patterns, such as the nucleotide (ATP/GTP) binding site, the Zinc fingers, the EGF domain, and the ferredoxin iron-sulfur domain. We believe some of these over-represented patterns were once part of the disabled protein coding genes containing the particular protein motif. As indicated by the very low *p*-value for the top-ranked patterns, there is almost zero chance that the enrichment of these patterns could have occurred just by chance. The Tyrosine kinase phosphorylation site (TYR_PHOSPHO_SITE) and RNA-binding region (RRM_RNP_1) are the two most under-represented patterns in both the fly and the worm intergenic regions. A possible explanation for the occurrences of these under-represented patterns is that the nucleotide sequences corresponding to some PROSITE patterns are disfavored by the chromosomal DNA structure.

For each pattern in PROSITE, the authors of the database assign a biological role, such as "Post-translational Modification" or "Domains," as shown the rightmost column in Table 3. We investigated whether there is correlation between the pattern occurrences and their biological

roles. For each genome we examined the functional annotations of over- and under-represented patterns in Table 3. Overall, the functional categories are fairly evenly distributed amongst the over-represented patterns. Nevertheless, there is a slight enrichment for transcription factors (e.g. ZINC_FINGER_C2H2_1, ZINC_FINGER_C3HC4, BZIP_BASIC and LEUCINE_ZIPPER) and several Cysteine-rich motifs (e.g. 4FE4S_FERREDOXIN, EGF_1, EGF_2, and PA2_HIS). It is known that worm has many pseudogenes for 7-tm transmembrane receptors [9]; no conserved protein motifs from this protein family are among the top over-represented motifs in worm genome.

Comparison Across Genomes: joint occurrence of a pattern in two genomes

It is interesting to compare the protein pattern occurrences across genomes, to see which patterns may be over- or under-represented in more than one species. Figure 1 shows two Venn diagrams comparing pattern occurrences between two triples of organisms: worm/fly/human and worm/fly/yeast. It is clear from the Venn diagrams that those patterns that are over-represented in one genome are likely to be over-represented in another. The worm and the fly share 19 over-represented patterns between them. This cannot be a coincidence, considering that in each genome less than 5% of the total PROSITE patterns are over-represented. Such a trend can be seen in the under-represented patterns as well.

It is also obvious from the Venn diagrams that the worm and the fly share more over-represented pseudomotif patterns with the human than with yeast. One trivial explanation is that the yeast intergenic sequences are much shorter than those of the worm, fly or human; thus the closeness between the pattern occurrences merely reflects the closeness of the size of intergenic sequences. However, we believe that the similarity between the pattern occurrences in worm, fly and human also reflects that these three organisms are closer to each other evolutionarily than to the single-celled yeast.

Table 2(B) lists the distribution of the 1319 patterns according to their joint occurrence in both fly and worm genomes. That is, it lists how many times specific PROSITE patterns occur similarly or differently in both these genomes. For instance, the upper-left corner of the table shows that 19 patterns are over-represented in both organisms and the upper-right shows that there are no patterns that are over-represented in the worm and under-represented in the fly. Overall, there is a consistency between the two genomes as evident from the larger values of the "on-diagonal" rather than "off-diagonal" entries. That is, the majority of pseudomotif patterns over-represented in the worm are also over-represented in the fly, and the same is true for the under-represented patterns.

Fly Has More Over-represented Patterns Though It Has Fewer Pseudogenes

Overall, we find more over-represented pseudomotif patterns in the fly genome than in the worm genome. This is somewhat surprising since the fly has considerably fewer pseudogenes than the worm, only ~200 compared with over 2000 in the worm [9, 17, 18]. Although we have masked out all the annotated repetitive elements, we were still concerned that there may be some fragments of the repetitive elements left out in the fly genome which may have introduced artifacts in our study. We translated in six-frame the sequences of the most common *Drosophila* transposable elements (TE), *copi*a and *gypsy*, and searched for PROSITE motifs in the predicted protein sequences. Among the PROSITE motifs identified in the transposable elements, only MYRISTYL (PS00008) and LEUCINE_ZIPPER (PS00029) are over-represented in the whole fly genome (Table 3). Thus the over-representation of the pseudomotifs in fly genome did not originate from the overlooked transposable elements.

Petrov and colleagues [19, 20] have found that in the fly genome, deletions are about three times more frequent and eight times longer than those in mammals, leading to an approximate 24-fold increase in the rate of spontaneous DNA loss. These researchers concluded that such a high rate of DNA loss results in the scarcity of pseudogenes in the fly genome. There has been some

limited data on the rate of DNA loss in worm genome [21], but it is probably still inconclusive how it compares with the rate in fly. However, if we assume that the fly genome has a higher deletion rate than the worm and take this as a working hypothesis, then we can explain our puzzling finding that the fly has more over-represented patterns than the worm.

As implied by the term "rate of gene loss", a gene is "lost" when the entire protein coding sequence is no longer recognizable by either gene-finding or homology-search techniques as result of natural substitution, insertion or deletion. However, small fragments of the protein coding sequence can still stay intact during the course of evolution. Because these fragments have much shorter lengths than ordinary open reading frames, conventional homology-search techniques will miss them and fail to count them as pseudogenes. Nevertheless these short protein motifs can still be recovered by our pattern-matching procedure. The schematic diagram in Figure 2 shows the different evolutionary paths for disabled genes in the worm and fly genomes. If we assume that during the course of evolution relatively equal numbers of genes become disabled in both organisms, then because of a slower deletion rate, the disabled genes in the worm will accumulate fewer deletions than those in the fly, and thus will be more likely to be detected as pseudogenes (and get masked out from our analysis). In the fly, in contrast, disabled genes are less likely to be recognized as pseudogenes, due to a greater deletion rate; consequently they will be counted as pseudomotifs. Thus, it is not surprising that more pseudomotifs are over-represented in the fly than in the worm. We emphasize that our idea is based on the assumption that the fly genome in general has a higher deletion rate than the worm, an assumption that still needs to be confirmed experimentally or computationally.

Conclusion

The biological roles of the intergenic sequences in the genomes of higher organisms have puzzled scientists for many years. These sequences are sometimes called "junk DNA", since no biological function can be assigned to them. Here we analyzed the occurrences of conserved protein patterns in the intergenic region of several genomes. We found that 67 pseudomotif patterns are over-represented in the fly genome, and 34 in the worm. Among the over-represented patterns are the well-known Zinc finger, Leucine zipper, and nucleotide-binding motifs. We argue that the enrichment of these patterns in the intergenic region reflects the fact that some of them are remaining fragments of ancient disabled protein-coding genes. It is hard to estimate how many other protein fragments are hidden in the intergenic region and what percentage of the intergenic region used to code protein, since it is difficult to define what constitutes a protein feature. Nevertheless, what we have discovered here could shed new light on the origin and functions of intergenic sequences in higher organisms.

Additional proof: We also did a survey on potential trans-membrane helices in the intergenic region. Following Poisson modeling techniques similar to those described here, we also found more than the expected number of TM helices (manuscript in preparation).

Materials and Methods

Creating Intergenic Sequences

The raw genome sequences and the GFF annotation files for yeast (*Saccharomyces cerevisiae*) are downloaded from Saccharomyces Genome Database [22] (URL: http://genome-www.stanford.edu/Saccharomyces/). Those of worm (*Caenorhabditis elegans*) [1] are from Sanger Centre (URL: http://www.sanger.ac.uk); those of the fly (*Drosophila melanogaster*) are from Berkeley Drosophila Genome Project [17] (URL: http://www.fruitfly.org/); and those of human

(*Homo sapiens*) are from NCBI and the Sanger Centre [1, 2]. Pseudogene annotations for the yeast, worm and human genomes have been described previously [9-11]; pseudogenes for the fly will be reported elsewhere. Repeat sequences, missing nucleotides, introns and exons and pseudogenes are masked out from the sequence to produce fragments of intergenic sequences. Table 4 shows the statistics of these intergenic sequences for the four genomes. Note that the analyzed human intergenic sequence comprises only chromosomes 21 and 22, which make up about 3% of the whole human genome. We choose worm and fly as the focus of our across-species comparison because of the relatively small size of the yeast intergenic region (see Table 4) and also because of our human pseudogenes are only available for two chromosomes.

Scan For PROSITE Patterns

The intergenic nucleotide sequences are translated into amino acids (including stop codons) in six frames, and a PERL script is set up to automatically scan for short amino acid fragments that match a PROSITE pattern. The PROSITE database [13, 14] is a manually curated database of biologically significant sites, patterns and motifs; "the database is formulated in a computer-friendly way such that with appropriate computational tools it can be used to rapidly and reliably detect conserved protein patterns in sequences". There are a total of 1474 entries in PROSITE as of September 2000. They are classified as "rules", "patterns" or "profile matrices" according to the way each entry was constructed. Most of these entries are protein "patterns," which have lengths ranging from 3 to over 20 amino acids; only patterns longer or equal to 5 amino acids are used in this study. Table 3 shows examples of some of the patterns that figured prominently here. For a particular pattern, the translated intergenic sequences are scanned at every position progressively. If two overlapping matches for the same pattern are found in the same frame of translation, they are counted separately. The overlapping patterns are treated this way so the number of observed occurrences can be compared with the expected number of

matches calculated from a Poisson model (described below). For some of the "over-represented" patterns (see below), we also examined the translated amino acid sequences of the matches to confirm these matches are not the result of DNA sequence repetition.

Recently several other protein pattern databases similar to PROSITE have been developed, such as Bio-Dictionary [23, 24], BLOCKS [25, 26], PRINTS [27], EMOTIF [28, 29] and others. The reasons we chose to use PROSITE instead of these other databases are: (i) The PROSITE patterns are extracted and curated by human experts instead of being generated by automated alignment programs, and thus are more precise and contain fewer false-negatives. (ii) The patterns in PROSITE are represented in a format similar to regular expressions in Perl, while in other databases the patterns are represented by a group of amino acid sequence aligned together. (iii) PROSITE is a member of the InterPro consortium [30, 31], a comprehensive, integrated protein signature database.

Calculating Expectations By Poisson Approximation

The occurrence of a query pseudomotif pattern along a long target DNA sequence can be modeled as a Poisson process. We use L to denote the total number of amino acids in the translated intergenic sequences, and p to denote the probability that a pattern occurs by chance beginning at any position in the intergenic region. Thus the expected number of occurrences (or the Poisson mean) for this pattern is simply: $\lambda = L \bullet p$.

Strictly speaking, this only holds when the sequence being scanned is a continuous piece instead of a collection of sequence fragments. But the error introduced by our treatment is negligible because the number of fragments is relatively small compared with the total length of sequences (Table 4). The probability p is determined according to following formula:

$$p = \prod_{i=1}^{N} \left(\sum_{a \in S(i)} P_a \right) \qquad (1)$$

where N is the length of the pattern, i is the position index, S(i) represents the set of amino acids that are allowed at position i, and P_a represents the background frequency for amino acid a in the translated intergenic sequences. For example, for a pattern like "C-x-[DN]-x(4)-{FY}-M-C", the probability of a sequence of 10 amino acids matching such a pattern is :

$$p = P_C \bullet (1- P_{SC}) \bullet (P_D + P_N) \bullet (1- P_{SC})^4 \bullet (1- P_F - P_Y - P_{SC}) \bullet P_M \bullet P_C \qquad (2)$$

where the P_C, P_D, P_N, P_F, P_Y, P_M and P_{SC} are the background frequencies of amino acid Cys, Asp, Asn, Phe, Tyr, Met and stop codon. Note that the character x in the pattern formula indicates that any amino acid is allowed at a position, and 1- P_{SC} represents the sum of frequencies for 20 amino acids sans stop. For the background frequencies in equation (1), we chose to use the amino acid composition derived from translating the intergenic regions in six frames. This choice is not trivial, as we could use frequencies derived in many other ways. For instance, we did not use the amino-acid frequencies extracted from the proteome in each organism, since they do not contain stop codons; moreover, they are characteristics of protein coding regions instead of non-coding regions.

Another option is to first compute the tri-nucleotide frequencies from the intergenic mononucleotide composition and then derive amino-acid frequencies from these based on the genetic code. However, using the mononucleotide composition is not appropriate in our case since it would introduce errors in modeling as the result of the "genomic signature" phenomenon [32]. It has been known that for each organism, its genome has a characteristic "signature" defined as the ratio between the observed dinucleotide frequencies and the frequencies expected if neighboring nucleotides were chosen at random. Campbell et al. [33] defined "genomic signature profile" as an array $\{\rho^*_{XY} = f^*_{XY} / f^*_X f^*_Y\}$, where f^*_X, f^*_Y denote the frequency of the mononucleotide X and Y, and f^*_{XY} the frequency of the dinucleotide XY, both computed from the sequence concatenated with its inverted complement. For human, worm, fly and yeast, ρ^*_{XY} is not always close to 1 and can be as great as 1.29 and as small as 0.22 for some dinucleotide pairs.

Consequently, calculating tri-nucleotide frequencies directly from mononucleotide composition will introduce significant bias and will not reflect the real characteristics of the intergenic sequences.

Calculating Statistical Significance

It is expected that some of the pseudomotif pattern matches we found could have arisen by chance, so it is important to determine the statistical significance of the observed number of occurrences. For each PROSITE pattern the observed number of occurrences N from scanning the genome is compared with λ, the expected number computed from the Poisson model. Significance *p*-values for N are calculated according to the Poisson distribution to indicate the probability that the pattern could occur by chance N times or more in the intergenic region.

$$P(y \geq \mathrm{N}) = \sum_{y=\mathrm{N}}^{\infty} \frac{\lambda^{y}}{y!} e^{-\lambda} \qquad (3)$$

The Patterns With Gaps

Some PROSITE patterns have a gap in the middle and thus do not have fixed lengths. For example, entry PS00820 (GLUCOAMYLASE) has a pattern: "[STN]-[GP]-x(1,2)-[DE]-x-W-E-E-x(2)-[GS]", where x(1,2) represents a gap of one or two amino acids. To compute the Poisson mean for such patterns, we can think of them as a union set of several individual patterns, each with fixed length. The Poisson mean or the expected number of occurrences of the original pattern is just the sum of the Poisson means of all the individual patterns. For the pattern we mentioned above, the Poisson mean can be computed as $\lambda = \lambda_1 + \lambda_2$, where λ_1 and λ_2 are the Poisson means for the patterns "[STN]-[GP]-x(1)-[DE]-x-W-E-E-x(2)-[GS]" and "[STN]-[GP]-x(2)-[DE]-x-W-E-E-x(2)-[GS]," respectively.

Simulations on Random Amino Acid Sequences

We did simulations to verify that the general assumptions behind the Poisson model were practically satisfied here. A basic assumption behind the model is that each match (i.e. the rare event model by the Poisson process) occurs independently in sequence. This assumption does not strictly hold in the pattern matching process. Suppose we are searching for a pattern of length l. We compare it to a sequence window of length l starting at position n in the intergenic region, denoted as S(n, n+l-1). As we slide the window progressively by one amino acid, the sequence in the new window S(n+1, n+l) is overlapped with the sequence in the previous window S(n, n+l-1) at all positions except the ends. Thus, the probabilities for sequences in consecutive windows to match a pattern are not completely independent of each other. This could potentially introduce a bias into our Poisson model; however, it is minimal because the chance that a stretch of sequence matching a typical pattern is very small (typically << 1.0e-4). To verify the validity of our Poisson model, we ran 600 simulations on randomized amino acid sequences of length 2,000,000. The results show great agreement between simulation and the model, indicating that the way we compute the expectation is fairly accurate.

For those pseudomotifs that we found to have significantly more occurrences than expectation, we also conducted simulations to verify the accuracy of the expectation calculated based on our Poisson model. For these simulations, we shuffled translated amino acid sequences from intergenic sequences to create 100 randomized copies and then performed the same pattern searching procedures on these. The observed occurrences on these shuffled sequences are very close to the ones calculated from our Poisson model.

Website

We made available a website with detailed statistics of our pseudomotif pattern occurrences at http://bioinfo.mbb.yale.edu/genome/pseudogene/motif/ and http://genecensus.org/pseudogene/motif/

Acknowledgements

We thank Haiyuan Yu for creating the web interface for this paper. MG acknowledges NIH Structural Genomics Program and the Keck Foundation for financial support. ZZ thanks Drs. Jiang Qian and Nick Luscombe for helpful discussions.

References

1. International Human Genome Sequencing Consortium (2001) Initial sequencing and analysis of the human genome. *Nature,* **409**, 860-921.
2. Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G. *et al.* (2001) The sequence of the human genome. *Science,* **291**, 1304-1351.
3. Jareborg, N., Birney, E. & Durbin, R. (1999) Comparative analysis of noncoding regions of 77 orthologous mouse and human gene pairs. *Genome Res.* **9**, 815-824.
4. Wasserman, W. W., Palumbo, M., Thompson, W., Fickett, J. W. & Lawrence, C. E. (2000) Human-mouse genome comparisons to locate regulatory sites. *Nature Genet.* **26**, 225-228.
5. Shabalina, S. A. & Kondrashov, A. S. (1999) Pattern of selective constraint in *C. elegans* and *C. briggsae* genomes. *Genet. Res.* **74**, 23-30.
6. Bergman, C. M. & Kreitman, M. (2001) Analysis of conserved noncoding DNA in *Drosophila* reveals similar constraints in intergenic and intronic sequences. *Genome Res.* **11**, 1335-1345.

7. Mighell, A. J., Smith, N. R., Robinson, P. A. & Markham, A. F. (2000) Vertebrate pseudogenes. *FEBS Lett.* **468**, 109-114.

8. Vanin, E. F. (1985) Processed pseudogenes: characteristics and evolution. *Annu. Rev. Genet.* **19**, 253-272.

9. Harrison, P. M., Echols, N. & Gerstein, M. B. (2001) Digging for dead genes: an analysis of the characteristics of the pseudogene population in the *Caenorhabditis elegans* genome. *Nucleic Acids Res.* **29**, 818-830.

10. Harrison, P., Kumar, A., Lan, N., Echols, N., Snyder, M. & Gerstein, M. (2002) A small reservoir of disabled ORFs in the yeast genome and its implications for the dynamics of proteome evolution. *J. Mol. Biol.* **316**, 409-419.

11. Harrison, P., Hegyi, H., Balasubramanian, S., Luscombe, N., Bertone, P., Echols, N. *et al.* (2002) Molecular fossils in the human genome: identification and analysis of the pseudogenes in chromosomes 21 and 22. *Genome Res.* **12**, 272-280.

12. Harrison, P. M. & Gerstein, M. (2002) Studying genomes through the aeons: protein families, pseudogenes and proteome evolution. *J. Mol. Biol.* **318**, 1155-1174.

13. Hofmann, K., Bucher, P., Falquet, L. & Bairoch, A. (1999) The PROSITE database, its status in 1999. *Nucleic Acids Res.* **27**, 215-219.

14. Bucher, P. & Bairoch, A. (1994) A generalized profile syntax for biomolecular sequence motifs and its function in automatic sequence interpretation. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* **2**, 53-61.

15. Kasuya, A. & Thornton, J. M. (1999) Three-dimensional structure analysis of PROSITE patterns. *J. Mol. Biol.* **286**, 1673-1691.

16. Gerstein, M. (1998) How representative are the known structures of the proteins in a complete genome? A comprehensive structural census. *Folding & Design* **3**, 497-512.

17. Adams, M. D., Celniker, S. E., Holt, R. A., Evans, C. A., Gocayne, J. D., Amanatides, P. G. *et al.* (2000) The genome sequence of *Drosophila melanogaster*. *Science,* **287**, 2185-2195.

18. Harrison, P. M., Kumar, A., Lang, N., Snyder, M. & Gerstein, M. (2002) A question of size: the eukaryotic proteome and the problems in defining it. *Nucleic Acids Res.* **30**, 1083-1090.

19. Petrov, D. A., Lozovskaya, E. R. & Hartl, D. L. (1996) High intrinsic rate of DNA loss in Drosophila. *Nature,* **384**, 346-349.

20. Petrov, D.A. & Hartl, D.L. (1998) High rate of DNA loss in the *Drosophila melanogaster* and *Drosophila virilis* species groups. *Mol. Biol. Evol.* **15**, 293-302.

21. Robertson, H. M. (2000) The large *srh* family of chemoreceptor genes in *Caenorhabditis nematodes* reveals processes of genome evolution involving large duplications and deletions and intron gains and losses. *Genome Res.* **10**, 192-203.

22. Cherry, J. M., Adler, C., Ball, C., Chervitz, S. A., Dwight, S. S., Hester, E. T. *et al.* (1998) SGD: *Saccharomyces* Genome Database. *Nucleic Acids Res.* **26**, 73-79.

23. Rigoutsos, I., Floratos, A., Ouzounis, C., Gao, Y. & Parida, L. (1999) Dictionary building via unsupervised hierarchical motif discovery in the sequence space of natural proteins. *Proteins,* **37**, 264-277.

24. Rigoutsos, I., Gao, Y., Floratos, A. & Parida, L. (1999) Building dictionaries of 1D and 3D motifs by mining the Unaligned 1D sequences of 17 archaeal and bacterial genomes. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, 223-233.

25. Henikoff, J. G., Greene, E. A., Pietrokovski, S. & Henikoff, S. (2000) Increased coverage of protein families with the blocks database servers. *Nucleic Acids Res.* **28**, 228-230.

26. Henikoff, S., Henikoff, J. G. & Pietrokovski, S. (1999) Blocks+: a non-redundant database of protein alignment blocks derived from multiple compilations. *Bioinformatics,* **15**, 471-479.

27. Attwood, T. K., Croning, M. D., Flower, D. R., Lewis, A. P., Mabey, J. E., Scordis, P. *et al.* (2000) PRINTS-S: the database formerly known as PRINTS. *Nucleic Acids Res.* **28**, 225-227.

28. Huang, J. Y. & Brutlag, D. L. (2001) The EMOTIF database. *Nucleic Acids Res.* **29**, 202-204.

29. Nevill-Manning, C. G., Wu, T. D. & Brutlag, D. L. (1998) Highly specific protein sequence motifs for genome analysis. *Proc. Natl Acad. Sci. USA,* **95**, 5865-5871.

30. Apweiler, R., Attwood, T. K., Bairoch, A., Bateman, A., Birney, E., Biswas, M. *et al.* (2001) The InterPro database, an integrated documentation resource for protein families, domains and functional sites. *Nucleic Acids Res.* **29**, 37-40.

31. Apweiler, R., Attwood, T. K., Bairoch, A., Bateman, A., Birney, E., Biswas, M. *et al.* (2000) InterPro--an integrated documentation resource for protein families, domains and functional sites. *Bioinformatics,* **16**, 1145-1150.

32. Gentles, A. J. & Karlin, S. (2001) Genome-scale compositional comparisons in eukaryotes. *Genome Res.* **11**, 540-546.

33. Campbell, A., Mrazek, J. & Karlin, S. (1999) Genome signature comparisons among prokaryote, plasmid, and mitochondrial DNA. *Proc. Natl Acad. Sci. USA,* **96**, 9184-9189.

Figure Legends

Figure 1. Venn diagrams showing the co-occurrences of over-represented PROSITE patterns in the intergenic region of the worm, fly and yeast genomes (top) and the worm and yeast genomes and the human chromosome 21, 22 (bottom).

Figure 2. Schematic diagram showing that as the result of a high deletion rate and greater deletion length, disabled genes in the fly genome are likely to be left as smaller gene pieces (pseudomotifs). However, in the worm genome, disabled genes are more likely to be recognized pseudogenes.

Table 1. Occurrences of pseudomotif patterns in the intergenic sequences of (A) fly genome and (B) worm genome

(A)

Rank (fly)	Rank (worm)	Pattern ID	Occurrence / Expectation	*p*-value
1	**12**	MYRISTYL	9.0e+5/8.4e+5	1.0e-300
2	**2**	ATP_GTP_A	5392/3600	3.3e-173
3	**11**	TONB_DEPENDENT_REC_1	1.5e+5/1.2e+4	5.6e-158
4	**7**	ZINC_FINGER_C2H2_1	449/230	3.0e-42
5	**26**	4FE4S_FERREDOXIN	89/19	2.2e-31
6	826	LDLRA_1	9/0.008	3.7e-25
7	**16**	EGF_1	462/180	6.7e-16
8	126	BZIP_BASIC	98/36	5.6e-16
9	**14**	EGF_2	456/190	6.7e-16
10	**15**	PA2_HIS	344/160	8.9e-16
11	923	CUTICLE	5/0.005	2.0e-20
12	**4**	LEUCINE_ZIPPER	6450/4800	7.0e-11
13	**9**	ATPASE_ALPHA_BETA	196/120	1.3e-10
14	215	SBP_BACTERIAL_3	26/5.4	1.5e-10
15	**8**	GATASE_TYPE_II	2.3e+4/2.1e+4	6.3e-10
16	409	TUBULIN	10/0.63	1.5e-9
17	**10**	LIPOCALIN	202/130	2.0e-9
18	40	CHAPERONINS_CPN60	8/0.4	1.1e-8
19	1227	ASX_HYDROXYL	63/29	3.1e-8
20	51	TNFR_NGFR_1	20/4.3	3.3e-8
...	…	....	....	....
(1315)	(1318)	RRM_RNP_1	469/600	1.60e-8
(1316)	1296	N6_MTASE	136/220	7.4e-10
(1317)	1303	ZINC_PROTEASE	118/210	3.2e-12
(1318)	1297	PROFLIN	1575/1900	8.7e-15
(1319)	(1319)	TYR_PHOSPHO_SITE	6.95e+4/8.0e+4	6.9e-316

(B)

Rank (worm)	Rank (fly)	Pattern ID	Occurrence / Expectation	*p*-value
1	229	ALDEHYDE_DEHYDR_GLU	96/20	1.6e-34
2	**2**	ATP_GTP_A	1841/1400	3.5e-30
3	187	SOD_CU_ZN_1	14/0.49	3.2e-16
4	**12**	LEUCINE_ZIPPER	3824/2700	1.8e-11
5	**45**	SOD_CU_ZN_2	5/0.021	3.3e-11
6	105	CECROPIN	30/7.1	1.0e-10
7	**4**	ZINC_FINGER_C2H2_1	113/59	2.8e-10
8	**15**	GATASE_TYPE_II	8508/6600	3.2e-10
9	**13**	ATPASE_ALPHA_BETA	90/45	2.3e-9
10	**17**	LIPOCALIN	83/44	1.0e-7
11	**3**	TONB_DEPENDENT_REC_1	5682/5300	1.1e-7
12	**1**	MYRISTYL	3.3e+5/3.1e+5	1.1e-6
13	**25**	ZINC_FINGER_C3HC4	17/4.4	3.9e-6
14	**9**	EGF_2	82/49	1.0e-5
15	**10**	PA2_HIS	68/40	3.5e-5
16	**7**	EGF_1	64/37	3.7e-5
17	**21**	INSULIN	18/5.8	2.8e-4
18	**27**	DNA_LIGASE_A1	39/21	4.7e-4
19	**60**	PROTEIN_KINASE_ATP	14/4.8	4.9e-4
20	1038	SASP_2	1/0	5.2e-4
...		....	....	....
(1315)	**64**	RNASE_PANCREATIC	25/48	2.0e-4
(1316)	(1307)	AA_TRNA_LIGASE_II_2	38/69	3.3e-5
(1317)	(1312)	CRYSTALLIN_BETAGAMMA	547/650	1.8e-5
(1318)	(1315)	RRM_RNP_1	231/400	3.0e-20
(1319)	(1328)	TYR_PHOSPHO_SITE	6.1e+4/6.9e+4	8.6e-177

The pseudomotif patterns are divided into three groups: the over-represented, the under-represented and the statistically non-significant. Over-represented patterns are those that have greater than expected number of occurrences in the genome and also have a *p*-value smaller than 0.01. The under-represented patterns have less than expected occurrences and also have a *p*-value smaller than 0.01. Statistically non-significant patterns are those patterns that have a *p*-value equal to or greater than 0.01 regardless of their actual occurrences relative to expectation. The over-represented patterns are placed at the top of the table and sorted by *p*-values in ascending order. Under-represented patterns are at the bottom of the table, and sorted by *p*-values in descending order. The statistically non-significant patterns are in the middle and sorted first by occurrences and then by *p*-value. By sorting the patterns this way, we can give each pattern a rank, from 1 to 1319, indicating "the level of enrichment". Only the top 20 and bottom 5 of the patterns are shown in the table. The columns in the table are ranks in each species, pattern accession number in PROSITE, pattern ID, the observed and expected number of occurrences, and the statistical significance *p*-value. In the first two columns, the rankings for the over-represented patterns in that particular organism are shown in bold face, the rankings for the statistically non-significant patterns are underlined, and the rankings for the under-represented patterns are enclosed in brackets. Detail on individual patterns can be found in Table 3.

Table 2.

(A) Summary of pseudomotif pattern occurrences in four genomes

	Yeast	**Worm**	**Fly**	**Human**
Number of over-represented patterns	6	34	67	21
Number of under-represented patterns	2	12	23	18
Non-significant (p-value ≥ 0.01)	1311	1273	1229	1280
Total: 1319 patterns from PROSITE database				

(B) Distributions of the pseudomotif patterns according to their occurrences in both the fly and worm genomes. For instance, the upper-left corner of the table shows that 19 patterns are over-represented in both the worm and fly genomes and the lower-right corner of the table shows that 7 patterns under-represented in bother genomes

Intersection		**Pattern occurrences in fly**		
		67 Over-represented	1229 Non-significant	23 Under-represented
Pattern occurrences in worm	34 Over-represented	19	15	0
	1273 Non-significant	47	1210	16
	12 Under-represented	1	4	7

Table 3. Examples of PROSITE patterns. This serves as a look-up table for details on the PROSITE patterns listed in Table 1 or elsewhere in the text. The four columns are pattern ID, pattern description, pattern formula, and pattern class (as described in the text). A brief explanation of pattern notations as quoted from the PROSITE is provided below. More details on the PROSITE website (URL: http://www.expasy.ch/prosite/)

- Each element in a pattern is separated from its neighbor by a "-".
- The symbol "x" is used for a position where any amino acid is accepted.
- Ambiguities are indicated by listing the acceptable amino acids for a given position, between brackets "[]".
- Ambiguities are also indicated by listing between a pair of braces "{}" the amino acids that are not accepted at a given position.
- Repetition of an element of the pattern is indicated by with a numerical value or a numerical range between parentheses following that element.

Pattern ID	Description	Pattern	Class
4FE4S_FERREDOXIN	4Fe-4S ferredoxin, iron-sulfur binding domain	C-x(2)-C-x(2)-C-x(3)-C-[PEG]	Electron transport proteins
AA_TRNA_LIGASE_II_1	Aminoacyl-transfer RNA synthetases class-II	[FYH]-R-x-[DE]-x(4,12)-[RH]-x(3)-F-x(3)-[DE]	Enzymes_Ligases
ALDEHYDE_DEHYDR_GLU	Aldehyde dehydrogenase family	[LIVMFGA]-E-[LIMSTAC]-[GS]-G-[KNLM]-[SADN]-[TAPFV]	Enzymes_Oxidoreductases
ASX_HYDROXYL	Aspartic acid and asparagine hydroxylation site	C-x-[DN]-x(4)-[FY]-x-C-x-C	Post-translational modifications
ATP_GTP_A	ATP/GTP-binding site motif A (P-loop)	[AG]-x(4)-G-K-[ST]	Domains
ATPASE_ALPHA_BETA	ATP synthase alpha and beta subunit, N-terminal	P-[SAP]-[LIV]-[DNH]-x(3)-S-x-S	Enzymes_Hydrolases
BZIP_BASIC	bZIP (Basic-leucine zipper) transcription factor family	[KR]-x(1,3)-[RKSAQ]-N-x(2)-[SAQ](2)-x-[RKTAENQ]-x-R-x-[RK]	DNA or RNA associated proteins
CECROPIN	Cecropin	W-x(0,2)-[KDN]-x(2)-K-[KRE]-[LI]-E-[RKN]	Hormones and active peptides

CHAPERONINS_CPN60	Chaperonin cpn60 (60Kd subunit)	A-[AS]-x-[DEQ]-E-x(4)-G-G-[GA]	Protein secretion and chaperones
CRYSTALLIN_BETAGAMMA	Crystallin	[LIVMFYWA]-x-{DEHRKSTP}-[FY]-[DEQHKY]-x(3)-[FY]-x-G-x(4)-[LIVMFCST]	Structural proteins
CUTICLE	Insect cuticle protein	G-x(7)-[DEN]-G-x(6)-[FY]-x-A-[DNG]-x(2,3)-G-[FY]-x-[AP]	Structural proteins
DNA_LIGASE_A1	ATP-dependent DNA ligase AMP-binding site	[EDQH]-x-K-x-[DN]-G-x-R-[GACIVM]	Enzymes_Ligases
EGF_1	EGF-like domain	C-x-C-x(5)-G-x(2)-C	Domains
EGF_2	EGF-like domain	C-x-C-x(2)-[GP]-[FYW]-x(4,8)-C	Domains
GATASE_TYPE_II	Glutamine amidotransferase class-II	<x(0,11)-C-[GS]-[IV]-[LIVMFYW]-[AG]	Enzymes_Transferases
INSULIN	Insulin/IGF/Relaxin family	C-C-{P}-x(2)-C-[STDNEKPI]-x(3)-[LIVMFS]-x(3)-C	Hormones and active peptides
LDLRA_1	Low density lipoprotein (LDL)-receptor class A (LDLRA) domain	C-[VILMA]-x(5)-C-[DNH]-x(3)-[DENQHT]-C-x(3,4)-[STADE]-[DEH]-[DE]-x(1,5)-C	Domains

LEUCINE_ZIPPER	Leucine zipper pattern	L-x(6)-L-x(6)-L-x(6)-L	DNA or RNA associated proteins
LIPOCALIN	Lipocalin-related protein and Bos/Can/Equ allergen	[DENG]-x-[DENQGSTARK]-x(0,2)-[DENQARK]-[LIVFY]-{CP}-G-{C}-W-[FYWLRH]-x-[LIVMTA]	Other transport proteins
MYRISTYL	N-myristoylation site	G-{EDRKHPFYW}-x(2)-[STAGCN]-{P}	Post-translational modifications
N6_MTASE	N-6 Adenine-specific DNA methylase	[LIVMAC]-[LIVFYWA]-x-[DN]-P-P-[FYW]	Enzymes_Transferases
PA2_HIS	Phospholipase A2	C-C-x(2)-H-x(2)-C	Enzymes_Hydrolases
PROFILIN	Profilin/allergen	<x(0,1)-[STA]-x(0,1)-W-[DENQH]-x-[YI]-x-[DEQ]	Structural proteins
PROTEIN_KINASE_ATP	Protein kinases ATP-binding region signature	[LIV]-G-{P}-G-{P}-[FYWMGSTNH]-[SGA]-{PW}-[LIVCAT]-{PD}-x-[GSTACLIVMFY]-x(5,18)-[LIVMFYWCSTAR]-[AIVP]-[LIVMFAGCKR]-K.	Enzymes_Transferases
RNASE_PANCREATIC	Pancreatic	C-K-x(2)-N-T-F	Enzymes_Hydrolases

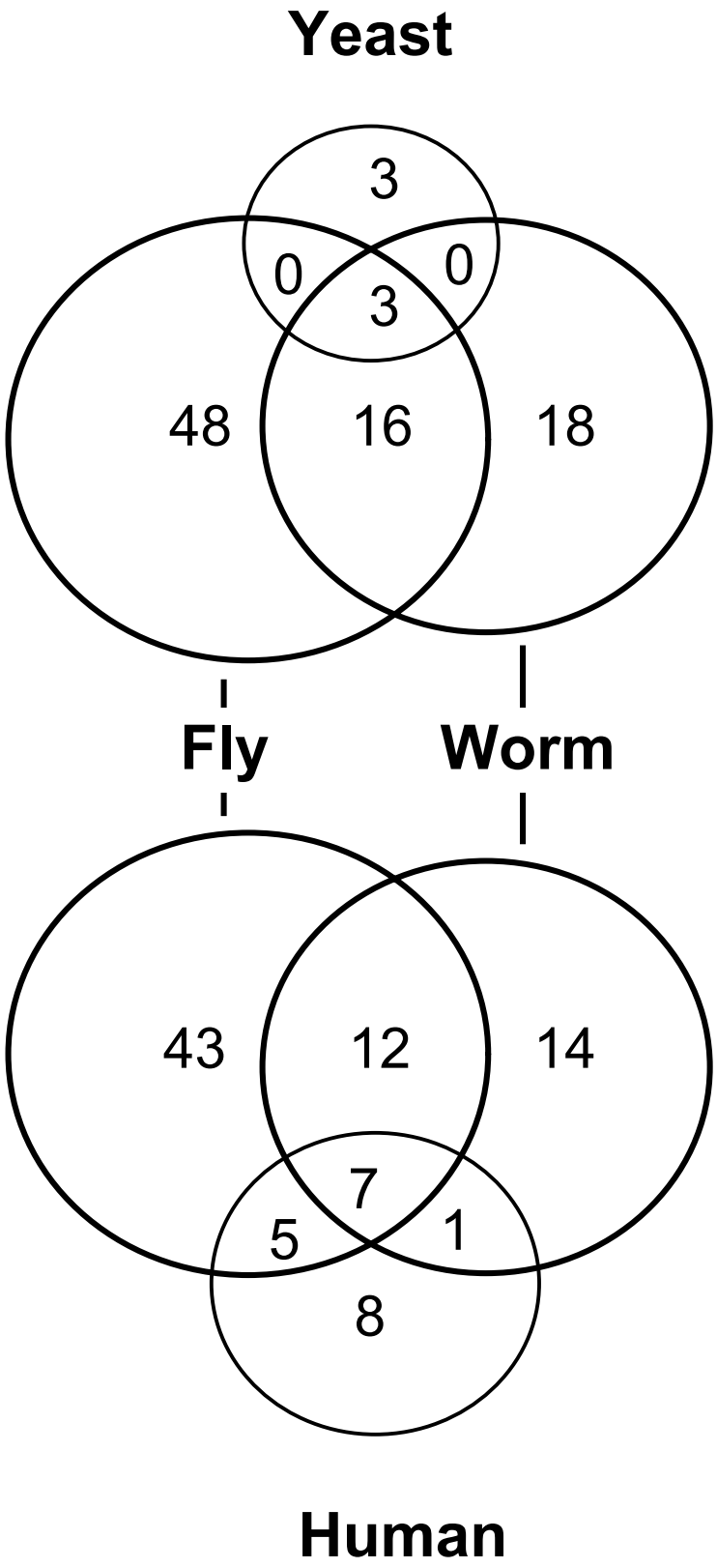
	ribonuclease family signature		
RRM_RNP_1	RNA-binding region RNP-1 (RNA recognition motif)	[RK]-G-{EDRKHPCG}-[AGSCI]-[FY]-[LIVA]-x-[FYLM]	DNA or RNA associated proteins
SASP_2	Small, acid-soluble spore proteins, alpha/beta type, signature 2	[KR]-[SAQ]-x-G-x-V-G-G-x-[LIVM]-x-[KR](2)-[LIVM](2)	DNA or RNA associated proteins
SBP_BACTERIAL_3	Bacterial extracellular solute-binding proteins, family 3	G-[FYIL]-[DE]-[LIVMT]-[DE]-[LIVMF]-x(3)-[LIVMA]-[VAGC]-x(2)-[LIVMAGN]	Other transport proteins
SOD_CU_ZN_1	Copper/Zinc superoxide dismutase	[GA]-[IMFAT]-H-[LIVF]-H-x(2)-[GP]-[SDG]-x-[STAGDE]	Enzymes_Oxidoreductases
SOD_CU_ZN_2	Copper/Zinc superoxide dismutase	G-[GN]-[SGA]-G-x-R-x-[SGA]-C-x(2)-[IV]	Enzymes_Oxidoreductases
TNFR_NGFR_1	TNFR/CD27/30/40/95 cysteine-rich region	C-x(4,6)-[FYH]-x(5,10)-C-x(0,2)-C-x(2,3)-C-x(7,11)-C-x(4,6)-[DNEQSKP]-x(2)-C	Receptors
TONB_DEPENDENT_REC_1	TonB-dependent	<x(10,115)-[DENF]-	Receptors

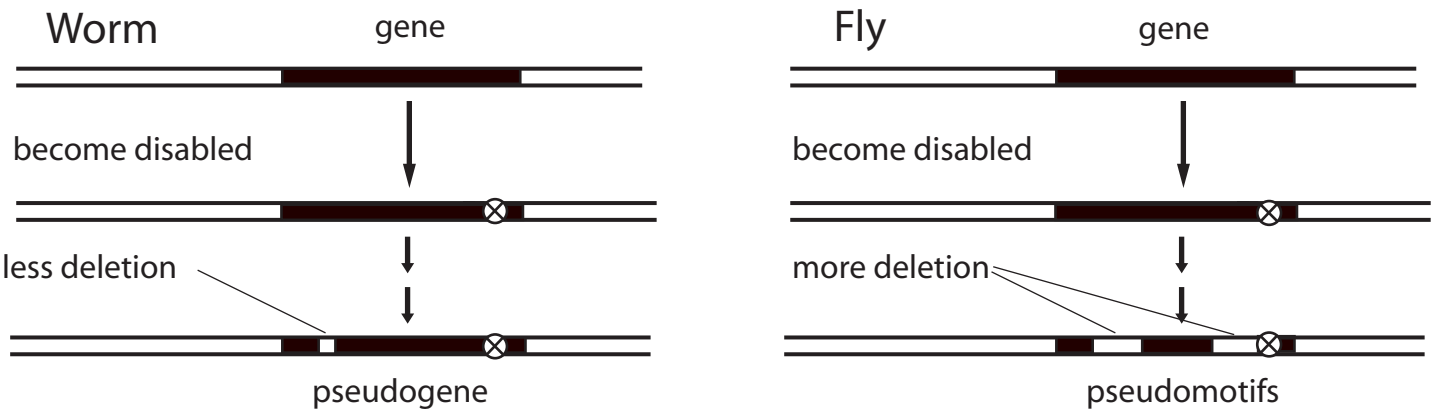
	receptor protein	[ST]-[LIVMF]-[LIVSTEQ]-V-x-[AGP]-[STANEQPK]	
TUBULIN	Tubulin family	[SAG]-G-G-T-G-[SA]-G	Structural proteins
TYR_PHOSPHO_SITE	Tyrosine kinase phosphorylation site	[RK]-x(2,3)-[DE]-x(2,3)-Y	Post-translational modifications
ZINC_FINGER_C2H2_1	Zinc finger, C2H2 type	C-x(2,4)-C-x(3)-[LIVMFYWC]-x(8)-H-x(3,5)-H	DNA or RNA associated proteins
ZINC_FINGER_C3HC4	RING finger	C-x-H-x-[LIVMFY]-C-x(2)-C-[LIVMYA]	DNA or RNA associated proteins
ZINC_PROTEASE	Neutral zinc metallopeptidases, zinc-binding region	[GSTALIVN]-x(2)-H-E-[LIVMFYW]-{DEHRKP}-H-x-[LIVMFYWGSPQ]	Enzymes_Hydrolases

Table 4. Statistics of intergenic sequences in four genomes

		Yeast	**Worm**	**Fly**	**Human ***
T	Total nucleotides in genome	12.06 M	100.09 M	116.12 M	68.56 M
I	Nucleotides in intergenic region	2.84 M	41.51 M	60.85 M	21.72 M
	Number of intergenic fragments	6,182	45,198	59,525	59,316
	Percentage (I/T)	23.5%	41.4%	52.4%	31.7%
P	Estimated nucleotides in "pseudomotifs"	0.079 M	1.28 M	2.08 M	0.71 M
	Percentage (P/I)	2.78%	3.08%	3.42%	3.27%

* Chromosome 21 and 22 only


Yeast
3
0
0
3
48
16
18
Fly
Worm
43
12
14
7
5
1
8
Human


Worm
gene
become disabled
less deletion
pseudogene
Fly
gene
become disabled
more deletion
pseudomotifs