A General Tendency for Conservation of Protein Length Across Eukaryotic Kingdoms

Daryi Wang,* Mufen Hsieh,* and Wen-Hsiung Li*†

*Computational and Evolutionary Genomics, Center for Genomics Research, Academia Sinica, Taipei, Taiwan; and †Department of Ecology and Evolution, University of Chicago

Protein elongation can occur in many ways, such as domain duplication or insertion and as recruitment of a transposable element fragment into the coding region, and it is believed to be a general tendency in protein evolution. Indeed, a previous study showed that yeast proteins are, on average, longer than their orthologs in bacteria, and in this study, we found that proteins in yeast, nematode, Drosophila, human, and Arabidopsis are, on average, longer than their orthologs in Escherichia coli. Surprisingly, however, we found conservation of protein sequence length across eukaryotic kingdoms. We collected 1,252 orthologous proteins from yeast, nematode, Drosophila, human, and Arabidopsis and found that the total length of these proteins is very similar among the five species and that there is no general tendency for a protein to increase or decrease in length. Furthermore, although paralogous proteins tend to undergo more sequence-length changes, there is also no general tendency for length increase. However, proteins that are commonly shared by Drosophila and human but not by yeast are, on average, substantially longer than proteins that are shared by yeast, Drosophila, and human. This is a puzzle that begs for an answer.

Introduction

It is commonly thought that proteins were originally short and simple but have become longer and more complex during evolution (see Li [1997]). Indeed, analyses of the human proteome have revealed a more complex domain organization of many human proteins in comparison to their homologous invertebrate and yeast proteins (Li et al. 2001; Venter et al. 2001; Kaesemann et al. 2002). An increase in protein length can occur by domain duplication or domain shuffling (see Li [1997], Kinch and Grishin [2002], and Ponting and Russell [2002]). In addition, fragments of transposable elements have been found in a large number of human proteins (Makalowski 2000; Lander et al. 2001; Li et al. 2001; Venter et al. 2001; Hughes and Coffin 2002). Finally, studies using elongation mutagenesis suggested that insertion of a random C-terminal tail could increase protein stability, and, in some cases, insertion of a peptide segment can improve the function of a protein (Matsuura et al. 1999; Chow et al. 2003; Claverie and Ogata 2003). Therefore, it seems that proteins would tend to increase in sequence length during evolution. Indeed, there is evidence that yeast proteins are, on average, longer than bacterial proteins (Zhang 2000).

On the other hand, an increase in peptide length may increase the energy cost of biosynthesis. A recent study of yeast genes suggested that natural selection favors shorter protein length for efficient synthesis (Akashi 2003). Studies using gene expression data showed strong correlation between codon usage bias and transcript abundance, suggesting that natural selection tends to increase the speed of protein synthesis (Moriyama and Powell 1998; Pal, Papp, and Hurst 2001; Akashi 2003). Also, it has been reported that proteins in the parasitic eukaryote Encephalitozoon cuniculi tend to be shorter than their eukaryotic orthologs (Katinka et al. 2001). In addition, spontaneous deletion tends to occur more often than spontaneous insertion in DNA sequences (de Jong and Ryden 1981). Thus, it is unclear whether in general proteins tend to increase in length.

Because there seems to be no study of protein sequence-length evolution in eukaryotes, we conducted a genome-wide comparison of protein lengths across eukaryotic kingdoms, taking advantage of the increasing abundance of genomic data in eukaryotes.

Materials and Methods

The protein sequence data were obtained from the following sources: E. coli (NCBI, ftp://ncbi.nlm.nih.gov/genomes/Bacteria/Escherichia_coli_K12/, 03/20/2003 version); a total of 4,225 protein sequences were obtained from the database. Yeast (Saccharomyces cerevisiae): (Saccharomyces Genome Database, ftp://gene-stanford.edu/pub/yeast/data_download/sequence/genomic_sequence/orf_protein/orf_dubious_trans, 03/08/04 version); 5,878 protein sequences were used. Nematode (Caenorhabditis elegans): (Sanger, ftp://ftp.sanger.ac.uk/pub/wormbase/current_release/ release 115); 22,227 protein sequences were used. Drosophila melanogaster: (Berkeley Drosophila Genome Project, http://www.fruitfly.org/cgi-bin/seq_tools/fasta_download.cgi, release 3); 18,498 protein sequences were used. Human (Homo sapiens): (European Bioinformatics Institute, http://www.ebi.ac.uk/proteome/index.html?http://www.ebi.ac.uk/proteome/HUMAN/), a nonredundant proteome set of SWISS-PROT, TrEMBL, and Ensembl entries; 28,657 protein sequences were obtained from the database. Arabidopsis thaliana: (Tigr, ftp://ftp.tigr.org/pub/data/a_thaliana/ath1/SEQUENCES/, release 4.0); 28,583 protein sequences were obtained from the database. The eukaryotic orthologous groups of proteins (KOGs) database was downloaded from NCBI (http://www.ncbi.nlm.nih.gov/COG/new/shokog.cgi); a total of 4,852 KOGs were obtained.

The commonly preserved (shared) proteins were defined by using yeast protein sequences as templates to BlastP against the protein sequences from the protein...
Table 1
Pairwise Wilcoxon Rank Sum Tests of the Lengths of Commonly Shared Proteins Between E. Coli and Each of Five Eukaryotes

<table>
<thead>
<tr>
<th>Species</th>
<th>Average Protein Length</th>
<th>Average Protein Length of E. coli</th>
<th>Pairs</th>
<th>P value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yeast</td>
<td>452.45</td>
<td>400.90</td>
<td>547</td>
<td>2.9×10⁻⁶</td>
</tr>
<tr>
<td>Nematode</td>
<td>461.02</td>
<td>415.75</td>
<td>590</td>
<td>1.4×10⁻⁴</td>
</tr>
<tr>
<td>Drosophila</td>
<td>455.38</td>
<td>399.85</td>
<td>610</td>
<td>2.1×10⁻⁶</td>
</tr>
<tr>
<td>Human</td>
<td>443.69</td>
<td>399.11</td>
<td>681</td>
<td>2.3×10⁻⁵</td>
</tr>
<tr>
<td>Arabidopsis</td>
<td>447.97</td>
<td>382.69</td>
<td>870</td>
<td>2.9×10⁻¹³</td>
</tr>
</tbody>
</table>

Note.—Criteria for the BlastP search: $E \leq e^{-10}$ and alignable length $\geq 50\%$ of the longer sequence. $P$ value: significance level of the Wilcoxon rank sum test.

databases of nematode, Drosophila, human, and Arabidopsis, separately. The orthologous proteins among the different species were defined by two criteria: the BlastP search with the E value $\leq e^{-10}$ and the alignable region between the two proteins greater than 50% of the longer protein. If more than one protein met the criteria, the protein with the highest hit score was chosen as the potential ortholog. The length of the multiple sequence alignment by the parsimony principle under the commonly accepted phylogeny of the five species in which the plant lineage branched off before the divergence between the yeast and animal lineages, and the insects (arthropods) are closer to humans than to nematode worms (the coelomata hypothesis) (see Blair et al. [2002], Hedges [2002] and Hughes and Friedman [2004]). An amino acid position was assumed to be in the ancestral sequence if it was present in at least three species or if the position was present in Arabidopsis and in at least one of the four other species. In this procedure, the Arabidopsis lineage has been given more weight because in the phylogeny used, it is the most divergent among the five eukaryotes studied. If this assumption is wrong and the yeast lineage is actually closer to the plant lineage than the animal lineage, it should have only a small effect on our inference because the three lineages are likely to be close to a trichotomy. Note that our inference makes no distinction between the coelomata hypothesis and the ecdyssoza hypothesis, which assumes that nematodes and insects form a clade and are equally distant to humans. Therefore, our inference holds under either hypothesis.

The KOGs database includes orthologous and paralogous genes of eukaryotic species. Each group is associated with a conserved and specific function (Tatusov et al. 2003). To compare the sequence lengths of orthologous and paralogous proteins, we used the 1,252 proteins commonly preserved among yeast, nematode, Drosophila, human, and Arabidopsis as queries to search the corresponding KOG databases (a total of 4,852 KOGs) and identified the paralogs of each set of orthologous proteins. Take the yeast as an example. We found 1,173 yeast proteins shared by the 1,252 proteins that have been commonly preserved among the five genomes and the KOGs database by identifying their accession numbers. These 1,173 proteins formed the set of the “orthologous” proteins that could be found in the KOGs database for the five genomes. Then, for each of the 1,173 proteins, we searched the KOGs database to find all other homologous yeast proteins and put them in the set of “paralogous proteins.” If there were n yeast paralogous proteins in the set of 1,173 proteins and there were other yeast paralogous proteins in the KOGs database, then all these proteins were clustered into n groups according to the rule of best hits in the BlastP search. If a group had more than one protein, then their average length was used in later analyses.

We also compared the lengths of the proteins commonly shared among yeast, Drosophila, and human with the proteins that were shared only by human and Drosophila but not yeast, which for simplicity are called derived proteins. The criteria for the BlastP search for commonly shared proteins were the same as described above.

Protein-length comparison was based on the lengths of polypeptides. Normality of the protein-length distribution was examined by the Kolmogorov-Smirnov (K-S) test; normality was rejected when the P value was smaller than 0.05. In addition, in view of the large variance of protein length, log10 transformation was applied to compress the range and to stabilize the variance, and then the K-S test was applied to test the normality. If the normality of either the original data or the log10-transformed data was accepted, the pairwise t-test was applied; otherwise, the pairwise Wilcoxon rank sum test was used.

Results
Longer Proteins in Eukaryotes Than in E. coli

Applying the BlastP search under the criteria described above, the commonly preserved proteins between E. coli and eukaryotes were collected. Here, we compare proteins from E. coli with their orthologous proteins in yeast (547 pairs), nematode (590 pairs), Drosophila (610 pairs), human (681 pairs), and Arabidopsis (870 pairs), separately, using the pairwise Wilcoxon rank sum test (table 1). The results show that the proteins in each of the five eukaryotic genomes studied are, on average, longer than their orthologous counterparts in E. coli. This tendency is the same as that in Zhang (2000), who compared yeast proteins with prokaryotic proteins.

Conservation of Protein Length among Eukaryotes

From the BlastP search, a total of 1,252 proteins commonly shared by yeast, nematode, Drosophila, human, and Arabidopsis were collected and analyzed. Figure 1 shows that the lengths of these 1,252 proteins are scattered as a right-skewed distribution, and the K-S test shows a departure from normality in the tails ($P < 1.16\times10^{-6}$).
The total numbers of amino acids in the 1,252 proteins for the five species are 547,044, 538,131, 548,927, 545,027, and 541,636, respectively. Human is actually similar to the four other species in the total length of these commonly shared proteins. In view of the large length variation among proteins, the protein lengths were log10 transformed, but normality still did not hold. Therefore, pairwise Wilcoxon rank sum tests were applied to compare the lengths. Figure 2 shows that yeast, nematode, Drosophila, human, and Arabidopsis have similar average lengths in the commonly shared proteins; all P values of pairwise Wilcoxon rank sum tests are large (>0.30), showing no significant difference between any two distributions.

Figure 3 shows more detailed comparisons of the human proteins with the ancestral proteins of yeast, nematode, Drosophila, human, and Arabidopsis. The length differences between human and ancestral protein sequences are nearly normally distributed around 0, indicating that there is a nearly equal probability for a protein to increase or decrease in length. Note that 48.7% of the 1,252 human proteins have not changed or have changed less than ±10 amino acids since their separation from the ancestral proteins (fig. 3a) and that 68.4% of these human proteins differ from their ancestral counterparts by less than ±5% of the ancestral sequence lengths (fig. 3b). The same patterns are found for the 1,252 proteins in yeast, nematode, Drosophila, and Arabidopsis (data not shown). These observations suggest that there is no general tendency for eukaryotic proteins to increase in sequence length.

Comparison of Paralogous Proteins

After separating the 4,852 KOGs database into orthologous and paralogous groups, we found 296 pairs (each orthologous versus paralogous from one KOG was counted as a pair) in yeast, 298 pairs in nematode, 287 pairs in Drosophila, 598 pairs in human, and 730 pairs in Arabidopsis (table 2). All KOG proteins that did not match any of the 1,252 proteins commonly preserved among the five eukaryotes or contain no paralog were excluded. The length ratios of paralogs to orthologs are shown in table 2. The pairwise Wilcoxon rank sum tests show that there is no significant difference in protein lengths between the paralogs and orthologs in the five eukaryote species studied, except for human. Human paralogous proteins are significantly shorter than their orthologous proteins by 14% (table 2), so paralogous proteins tend to be shorter than orthologous or ancestral proteins.

However, although, like orthologs, paralogs show no clear tendency to increase or decrease sequence length,
they have more frequent and larger length changes than do orthologs. This can be seen from a comparison of figure 4 with figure 3. For example, in figure 4, the proportion of proteins that have changed more than 10 amino acids is approximately 80% instead of approximately 50%, and the proportion of proteins that have changed more than 5% of their length is approximately 76% instead of approximately 30%. Clearly, the paralogs are more variable in length than the orthologs. Figure 4 shows only the data from human, but the same conclusion holds for data from each of the four other eukaryotic genomes.

Comparison of Young and Old Proteins

From our BlastP search, we found 2,376 proteins commonly shared by yeast, Drosophila, and human, which are defined as “old” proteins, and 4,077 proteins commonly shared by only Drosophila and human but not yeast, which are defined as “young” or “derived” proteins. (Note that a “young” protein defined here may actually not be a young protein, but represents a gene loss in the yeast lineage. However, the number of such cases is likely small compared with the total number [4,077]; for example, Krylov et al. [2003] inferred fewer than 500 gene losses in two fungus lineages. Another possibility is that a young protein here might actually be present also in the yeast proteome but was missed in the BlastP search because of sequence divergence.) Figure 5 shows that the 2,376 old proteins of yeast, Drosophila, and human are similar in length for the three species, and the 4,077 young proteins shared by Drosophila and human are also similar in length for the two species. These results are consistent with the above observation that protein length tends to be conserved among eukaryotes. When we compare the sizes of old (commonly preserved proteins) and young (newly derived proteins) proteins of Drosophila and human, the young proteins are significantly longer than old proteins by approximately 22% (table 3). That is, the newly derived proteins tend to be longer.

Discussion

In the present analysis of protein sequence length, we have not considered the possibility that an alternatively spliced form of a protein rather than the “wild type” form was used in comparison. This possibility tends to increase the variation in sequence length. Also, in some cases, a paralog rather than an ortholog might have been used, despite the use of stringent criteria in the search of orthologs among genomes. This possibility also tends to increase the variation in sequence length. For these reasons, the sequence-length variation we inferred could have been

<table>
<thead>
<tr>
<th>Species</th>
<th>Average Length of Orthologous Proteins</th>
<th>Average Length of Paralogous Proteins</th>
<th>Length Ratio of Paralogs/Orthologs</th>
<th>P Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yeast</td>
<td>439.28</td>
<td>469.36</td>
<td>1.07</td>
<td>0.34</td>
</tr>
<tr>
<td>Nematode</td>
<td>428.80</td>
<td>408.42</td>
<td>0.95</td>
<td>0.41</td>
</tr>
<tr>
<td>Drosophila</td>
<td>409.58</td>
<td>417.72</td>
<td>1.02</td>
<td>0.74</td>
</tr>
<tr>
<td>Human</td>
<td>432.72</td>
<td>366.75</td>
<td>0.87</td>
<td>1.28×10⁻⁵</td>
</tr>
<tr>
<td>Arabidopsis</td>
<td>411.95</td>
<td>397.85</td>
<td>0.97</td>
<td>0.20</td>
</tr>
</tbody>
</table>

Note.—Data were obtained from the KOGs database, which contains orthologous and paralogous proteins of each species. The paralogous proteins were identified by using the 1,252 commonly shared proteins as queries in this study. P value: significance of Wilcoxon rank sum test.
overestimated, which would strengthen our conclusion on the sequence-length conservation of orthologous proteins among eukaryotes. Note, however, that our methods selected only proteins that have been well conserved in sequence among the five genomes studied. For those orthologous proteins that have been missed by our methods, their length might have not been well conserved.

Previous studies have suggested that present-day proteins have gone through several stages of evolution and that an evolving polypeptide chain may grow in length by insertion of residues into the chain or to its tail (Lupas, Ponting, and Russell 2001; Aravind et al. 2002; Trifonov and Berezovsky 2003). Empirical studies also suggested that insertion of a peptide segment can occasionally increase protein stability or even improve function (Matsuura et al. 1999; Chow et al. 2003; Claverie and Ogata 2003). Consistently, our results show that the lengths of proteins in eukaryotes are, on average, longer than those of E. coli proteins, extending the result by Zhang (2000). Surprisingly, we found conservation of protein sequence length since the common ancestor of yeast, nematode, Drosophila, human, and Arabidopsis. A simple explanation for the above observations is that although in general protein length had increased during the evolution from prokaryotes to eukaryotes, the length seems to have been largely optimized in the common ancestor of fungi, animals, and plants.

It has been proposed that first protein domains evolved by recombination from a limited number of polypeptides (Soding and Lupas 2003), followed by genetic events such as point mutations, insertions, and deletions. Insertion (including internal duplication) expands the protein sequence, providing opportunities for additional function or functional improvement (Matsuura et al. 1999; Trifonov and Berezovsky 2003). A survey of genes in eukaryotes showed that internal duplications have occurred frequently in evolution, sometimes increasing the number of active sites and, thus, enhancing the protein function (see Li [1997]). However, our study showed no general tendency for eukaryotic proteins to increase length; indeed, the chance for a decrease in length is as high as that for an increase in length. Because for the 1,252 commonly preserved proteins, only 402 are essential (deletion lethal) genes in yeast, functional constraint may not be the primary factor for the conservation of sequence length. Rather, the conservation in sequence length is probably mainly caused by structural constraint.

Some studies have indeed pointed to the importance of structural constraint in protein evolution. For example, thermodynamic stability and folding kinetics were shown to exert pressure on the course of protein evolution (Dokholyan and Shakhnovich 2001). The limited diversity of protein domain was also suggested to be caused by structural constraint (Jones et al. 1998; Hou et al. 2003). Furthermore, the study of Yang, Gu, and Li (2003) on the relationship between protein dispensability and the rate of evolution suggested that structural constraints are more important in determining the rate of amino acid substitution in proteins than functional requirement. Finally, as is well known, protein three-dimensional structure studies have shown that protein structure is much better conserved than sequence (Ponting and Russell 2002; Soding and Lupas 2003). Our observation of sequence-length conservation during the evolution of eukaryotes may be explained by protein-structure conservation because changes in sequence length may often cause changes in structure.

We also note that although paralogous proteins often undergo deletions and insertions probably because of relaxation in functional constraint, they show no general tendency to increase sequence length. Also, although there seems to be a tendency for paralogous proteins to decrease in sequence length, the tendency is very weak because it is significant only in one (human) of the five species studied (fig. 4). This is somewhat surprising in view of the facts that spontaneous deletion occurs more often than spontaneous insertion and that a shorter protein requires a lower cost of biosynthesis.

Interestingly, “newly evolved” or “derived” proteins are, on average, substantially longer than “old” proteins (fig. 5). It is not clear how this has happened, but we speculate two possibilities. First, some of these proteins might have already been present in the common ancestor of yeast, Drosophila, and human but had gained insertions and underwent a period of rapid amino acid change. Second, some of these proteins could have been derived from duplicate genes that had undergone gene elongation and rapid sequence changes. In both cases, the proteins have undergone so much sequence change that they can no longer be detected by the BlastP search using yeast proteins as queries. In any case, many of the new proteins might have gained new function, partly through sequence elongation. Whether these speculations have any merit requires further investigation.

Acknowledgments

We thank Dr. Henry Horng-Shing Lu and Dr. Hung-Mo Sung for help in the analysis and George Zhang for comments. This study was supported by Academia Sinica, Taiwan, and by NIH grants (GM30998 and GM66104) to WHL.

Literature Cited


| table 3 | Wilcoxon Rank Sum Tests Length Differences Between Proteins Commonly Shared by Yeast, Drosophila, and Human and Derived Proteins Present Only in Drosophila and Human |
|---|---|---|---|
| | Average Length of | Average Length of | P value |
| | Common Proteins | Derived Proteins | |
| Drosophila | 405.80 | 513.43 | <2.2×10^{-16} |
| Human | 408.98 | 500.15 | 2.2×10^{-16} |

Note.—P value: significance of Wilcoxon rank sum test. Sample size: 2,376 common proteins; 4,077 derived proteins.

Takashi Gojobori, Associate Editor
Accepted September 8, 2004