BMC Evolutionary Biology

BioMed Central

Research article

The relationship of protein conservation and sequence length

David J Lipman*, Alexander Souvorov, Eugene V Koonin, Anna R Panchenko and Tatiana A Tatusova

Address: National Center for Biotechnology Information, National Institutes of Health, Bethesda, MD 20894, USA

E-mail: David J Lipman* - lipman@ncbi.nlm.nih.gov; Alexander Souvorov - souvorov@ncbi.nlm.nih.gov; Eugene V Koonin - koonin@ncbi.nlm.nih.gov; Anna R Panchenko - panch@ncbi.nlm.nih.gov; Tatiana A Tatusova - tatiana@ncbi.nlm.nih.gov
*Corresponding author

Published: 1 November 2002

BMC Evolutionary Biology 2002, **2**:20

Received: 6 September 2002
Accepted: 1 November 2002

This article is available from: http://www.biomedcentral.com/1471-2148/2/20

Abstract

Background: In general, the length of a protein sequence is determined by its function and the wide variance in the lengths of an organism's proteins reflects the diversity of specific functional roles for these proteins. However, additional evolutionary forces that affect the length of a protein may be revealed by studying the length distributions of proteins evolving under weaker functional constraints.

Results: We performed sequence comparisons to distinguish highly conserved and poorly conserved proteins from the bacterium *Escherichia coli*, the archaeon *Archaeoglobus fulgidus*, and the eukaryotes *Saccharomyces cerevisiae*, *Drosophila melanogaster*, and *Homo sapiens*. For all organisms studied, the conserved and nonconserved proteins have strikingly different length distributions. The conserved proteins are, on average, longer than the poorly conserved ones, and the length distributions for the poorly conserved proteins have a relatively narrow peak, in contrast to the conserved proteins whose lengths spread over a wider range of values. For the two prokaryotes studied, the poorly conserved proteins approximate the minimal length distribution expected for a diverse range of structural folds.

Conclusions: There is a relationship between protein conservation and sequence length. For all the organisms studied, there seems to be a significant evolutionary trend favoring shorter proteins in the absence of other, more specific functional constraints.

Introduction

Proteins evolve under a range of constraints. Probably the most studied constraints on proteins have to do with their specific function, for example, as enzymes, regulators or signaling molecules. In addition, more general constraints on protein evolution are apparent from studies showing a correlation between the base composition of a genome (i.e. GC content) and the overall amino acid composition of its proteins c.f. [1]. There can also be general constraints on protein length. For example, prokaryotes have shorter proteins on average than eukaryotes [2], and among the eukaryotes, the proteins of the microsporidium *Encephalitozoon cuniculi*, with an extremely compact genome, are smaller than the corresponding proteins in organisms with larger genomes [3].

The constraints due to specific protein function and these more general constraints might not always act in concert.

For example, Singer & Hickey [4] observed a weaker correlation between base composition and amino acid composition for conserved proteins as compared to rapidly evolving proteins. Thus, the more general constraints might be obscured for proteins, which evolve under intense specific functional constraints.

With this in mind, we studied the length distributions of rapidly and slowly evolving proteins from a range of organisms in an effort to detect general constraints on protein length.

Results and Discussion

To analyze the length distributions of proteins from a given organism, we start with a set of proteins from that organism chosen so as to minimize the number of partial proteins or sequences generated by *ab initio* gene prediction methods. We denote the subset of these proteins that share statistically significant similarity with proteins from organisms outside the given primary kingdom (Archaea, Eukaryota or Bacteria) the Conserved Set (i.e. slowly evolving proteins). We denote another subset the Nonconserved Set (i.e. rapidly evolving proteins) if they only match proteins from closely related organisms (e.g. human proteins to other mammals, or *Drosophila* proteins to other insects) or do not match proteins from any other organism (see Methods).

In addition to the length distributions of the Conserved and Nonconserved Set proteins, we also analyzed the length distribution of protein domains in a non-redundant set of protein structures derived from a range of eukaryotes and prokaryotes organisms, denoted the Minimal Structural Domain Set [5]. The Minimal Structural Domain Set contains 1882 domains defined purely on the basis of structural compactness. A chain is split between secondary structure elements whenever the ratio of intra- to inter-domain contacts exceeds a threshold [6]. This computational approach for determining domain boundaries splits multi-domain proteins into single domains and non-compact strands and loops are removed as well, thus even single-domain proteins may be shortened by this method. Given that the Minimal Structural Domain Set is derived from a non-redundant set of protein structures and that residues that are not compactly folded are removed, this set approximates the minimal length distribution possible for a diverse range of protein folds.

We computed the protein length histograms of the Conserved and Non-conserved Sets along with that of the Minimal Structural Domain Set for the bacterium *Escherichia coli* (Figure 1), the archaeon *Archaeoglobus fulgidus* (Figure 2), and the eukaryotes *Saccharomyces cerevisiae* (Figure 3), *Drosophila melanogaster* (Figure 4), and *Homo sapiens* (Figure 5). Note that in these figures, the numbers for the Minimal Structural Domain Set were scaled to the total number of proteins in the respective Nonconserved Set. Because annotation artifacts of genomic sequence are likely to be more frequent among the shorter proteins [7,8], we computed an additional length distribution for a set of *E. coli* proteins (Figure 1) that share statistically significant similarity to proteins from the closely related bacterium *Salmonella typhimurium* but not proteins from more distant species. This group of proteins, denoted the *Salmonella* Set, approximated the Nonconserved Set but avoided potential artifacts associated with spurious short "proteins" in the latter.

The most obvious observations coming from the comparison of the resulting distributions are:

- the Conserved Set proteins are, on average, longer than the Nonconserved Set proteins;

- the length distributions of the Nonconserved Sets have a relatively narrow peak, whereas those of the Conserved Sets are spread over a wider range of values;

- the histograms of the *Salmonella* Set and the Nonconserved Set from *E. coli* (Figure 1) have a similar shape and length range;

- the histograms for the Nonconserved Set proteins from *E. coli* and *A. fulgidus* match the histogram of the Minimal Structural Domain Set fairly closely;

- the peaks of the Nonconserved Set histograms from yeast and *Drosophila* have shifted slightly to the right (i.e. the proteins tend to be longer) compared to the Minimal Structural Domain Set and the peak for the human proteins has shifted still more to the right;

- the Conserved Set proteins from yeast, *Drosophila*, and human are, on average, longer than those from *E. coli* and *A. fulgidus*;

- the right shoulder of the Nonconserved Set histograms from yeast, *Drosophila*, and human also diverge more from the Minimal Structural Domain Set histogram than do the *E. coli* and *A. fulgidus* histograms.

To evaluate the sensitivity of these observations to the cutoff expectation value used for the sequence comparisons (see Methods), we varied this cutoff over a range of six orders of magnitude for the *E. coli* proteins. The above conclusions held for all cut-off values (Figure 6).

For purposes of clarity, Figures 1, 2, 3, 4, 5 only show histograms for the Conserved and Nonconserved Sets. However, one could generate a length distribution histogram

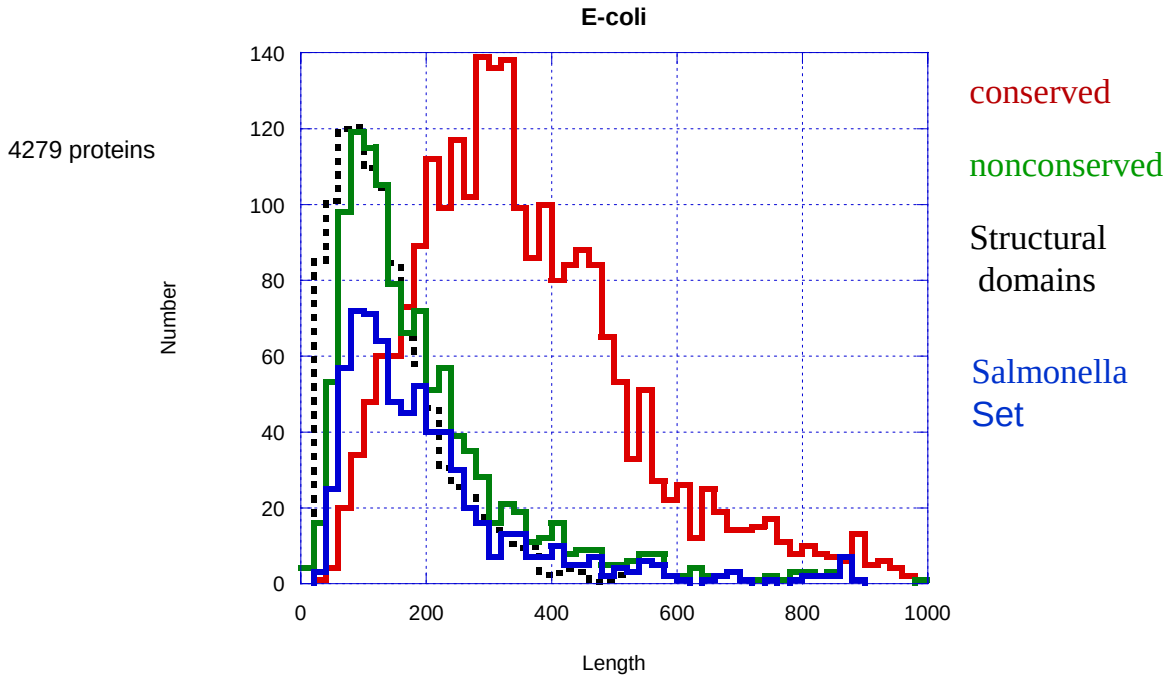

Figure 1
Protein Conservation versus Sequence Length – Escherichia coli Red curve is histogram of lengths of Conserved Set proteins, green curve is for Nonconserved Set, black dashed curve is normalized histogram of lengths of the Minimal Structural Domain Set, and blue curve is histogram of lengths of *Salmonella* Set proteins.

for proteins whose evolutionary rate is intermediate between these two extremes, denoted the Intermediate Set. Thus, for yeast, the Intermediate Set includes proteins that, the Intermediate Set histogram was positioned between the histograms for the Conserved and Nonconserved Sets (data not shown). In addition, similar results were obtained for a variety of eukaryotic organisms for which a representative sample of known full-length proteins were available.

Although Figures 1, 2, 3, 4, 5 show that, with increasing length, an increasing fraction of the proteins contain regions that are highly conserved, they do not indicate the fraction of residues that is conserved in these proteins. In Figure 7, we show the fraction of all conserved residues for all *E. coli* and *A. fulgidus* proteins (conserved and nonconserved) of varying lengths. With increasing length, a greater fraction of the residues are conserved, converging at approximately 80–90% for proteins greater than 400 residues long. Also in Figure 7, we show the contact density – the average number of contacts per amino acid residue – for known protein structures of varying lengths (see Methods). The curve for contact density shows good agreement with the curves corresponding to the fraction of conserved residues.

Discussion

The above results seem to indicate that conserved proteins are, in general, longer than non-conserved ones. It is highly unlikely that the above results are due to a detection bias given that these observations were unchanged when varying the cutoff expectation value used for the sequence comparisons (see Methods) between $E < 10^{-3}$ and $E < 10^{-9}$. A possible explanation for the insensitivity of these results to varying similarity thresholds is that, e.g., for the *E. coli* Conserved Set, 80% of the proteins had conserved re-

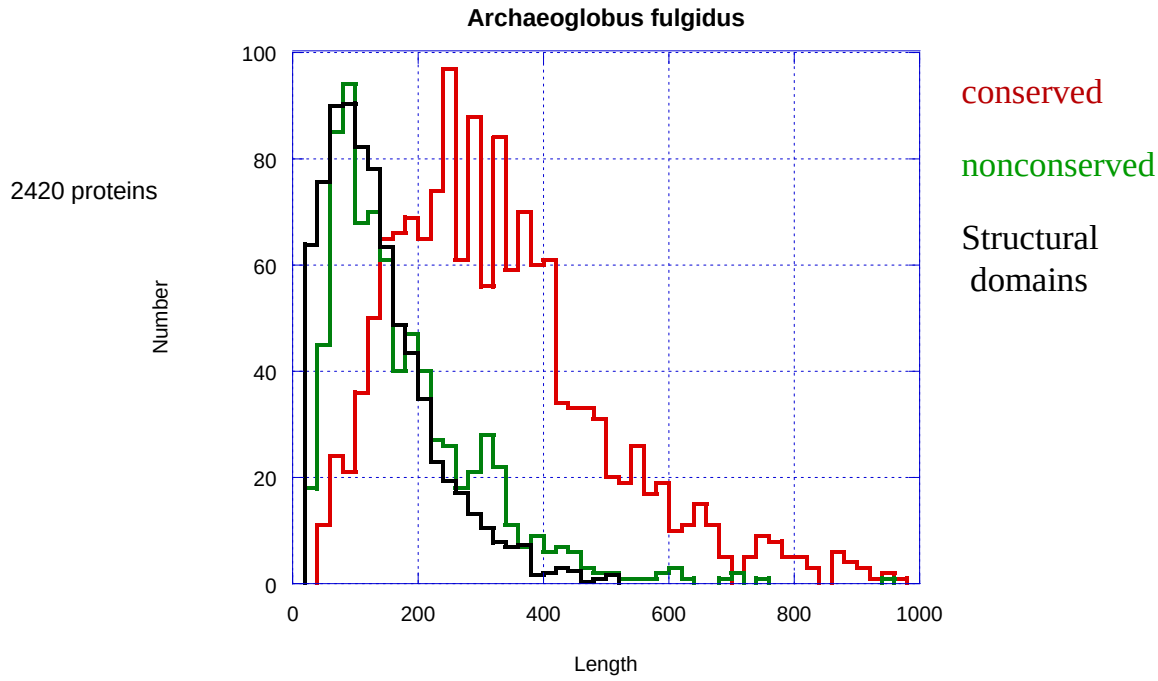

Figure 2
Protein Conservation versus Sequence Length – *Archaeoglobus fulgidus*. Red curve is histogram of lengths of Conserved Set proteins, green curve is for Nonconserved Set, black curve is normalized histogram of lengths of the Minimal Structural Domain Set.

gions (shared with protein from other kingdoms) over more than 75% of their length and thus would be easy to detect with most sequence comparison methods over a wide range of thresholds.

For the protein sequences determined by conceptual translation of genomic DNA, annotation artifacts would likely be more common among the shorter sequences and these would be classified into the Nonconserved Set. Thus, a possible explanation for the difference in the length distributions between the Nonconserved and Conserved Set proteins would be annotation artifacts for the proteins derived from genomic sequence. Skovgaard and colleagues [8] compared the length distribution of annotated microbial genome proteins matching known proteins with those that do not match a known protein. The sequences that did not have any matches were shorter and this was taken as evidence that too many short genes have been annotated in many genomes (i.e many of these short genes are artifacts). To test this possibility for the *Escherichia coli* proteins, we generated the length distribution for the *Salmonella* Set, a subset of *E. coli* proteins that match proteins from *Salmonella* but do not match proteins from more distant organisms. It is estimated that *Salmonella* and *E. coli* diverged about 100 million years ago [9] and thus a statistically significant similarity between sequences from these bacteria indicates that the corresponding genes evolve under purifying selection. Although this does not prove that all these genes encode proteins (i.e., some of them might encode heretofore uncharacterized regulatory RNAs), requiring a statistically significant similarity to *Salmonella* sequences greatly reduces the chance of retaining annotation artifacts. Although there are fewer proteins in the *Salmonella* Set, its length distribution is essentially the same as that of the Nonconserved Set (Figure 1).

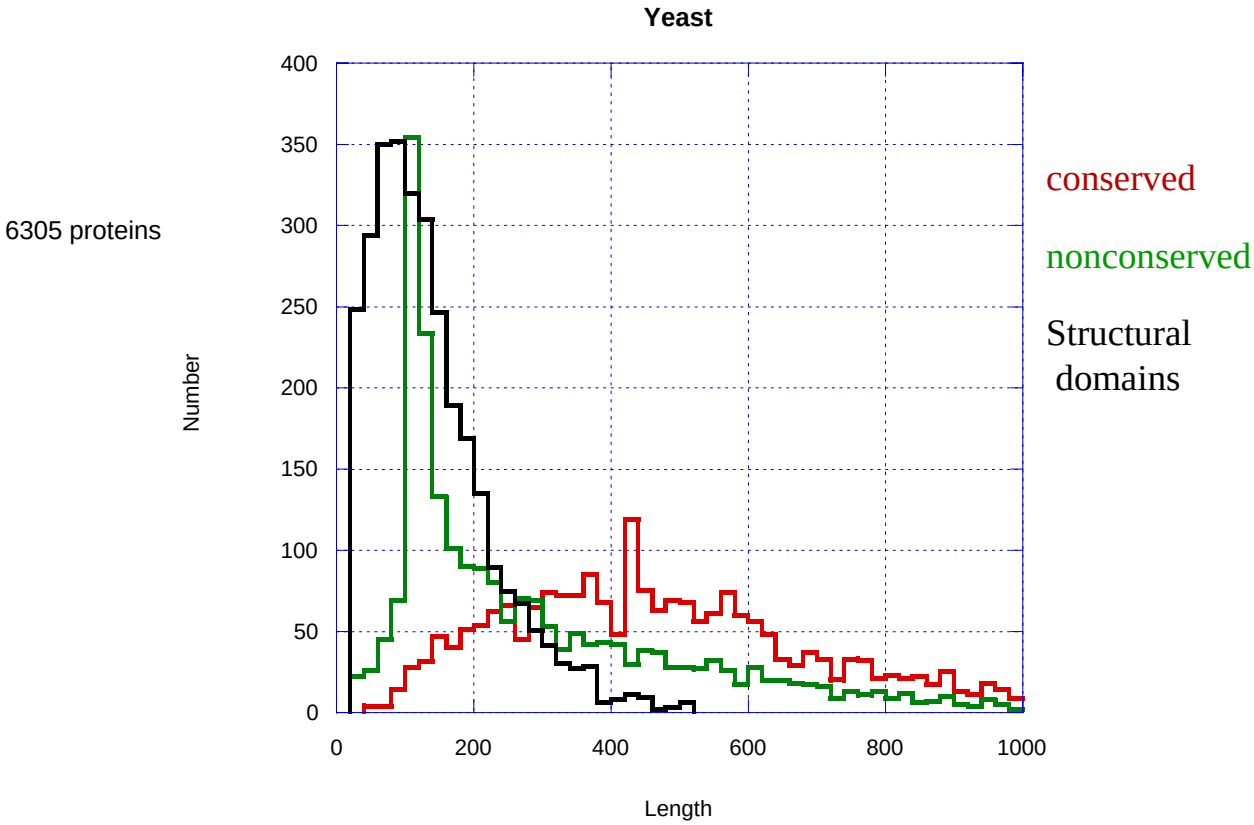

Figure 3
Protein Conservation versus Sequence Length – *Saccharomyces cerevisiae*

Furthermore, although it is likely that there is a greater fraction of annotation artifacts among the Nonconserved set proteins derived from genome annotations, this is unlikely to be true for the human and *Drosophila* proteins analyzed here because they have been derived from cDNA sequences. To further reduce the chance of annotation errors, for the *Drosophila* set, we avoided cDNA sequences generated from high throughput cDNA projects. Thus, annotation artifacts are unlikely to explain the results shown in Figures 1,2,3,4,5.

A challenging problem for biologists trying to make sense of genomic sequence, particularly for the eukaryotes, is that shorter proteins are more difficult to predict on purely statistical grounds [8] and are also less likely to have confirmatory homologies in other organisms. Thus, without functionally cloned cDNA transcripts, it becomes hard to distinguish artifacts from rapidly evolving genes and a conservative approach may result in under representation of the shorter eukaryotic proteins in the databases. Consistent with this possibility is the rightward shift of the Nonconserved Set proteins of the eukaryotes as compared to that of the prokaryotes.

One generally assumes that the length of a protein is largely determined by its functions. The relatively wide variance in sequence length of the members of the Conserved set reflects the diverse range of specific functional roles for these proteins. The Nonconserved set proteins however are, on average, shorter than the conserved proteins, with the poorly conserved *E. coli* and *A. fulgidus* proteins closely approximating the minimal length distribution possible for globular proteins, as represented by the Minimal

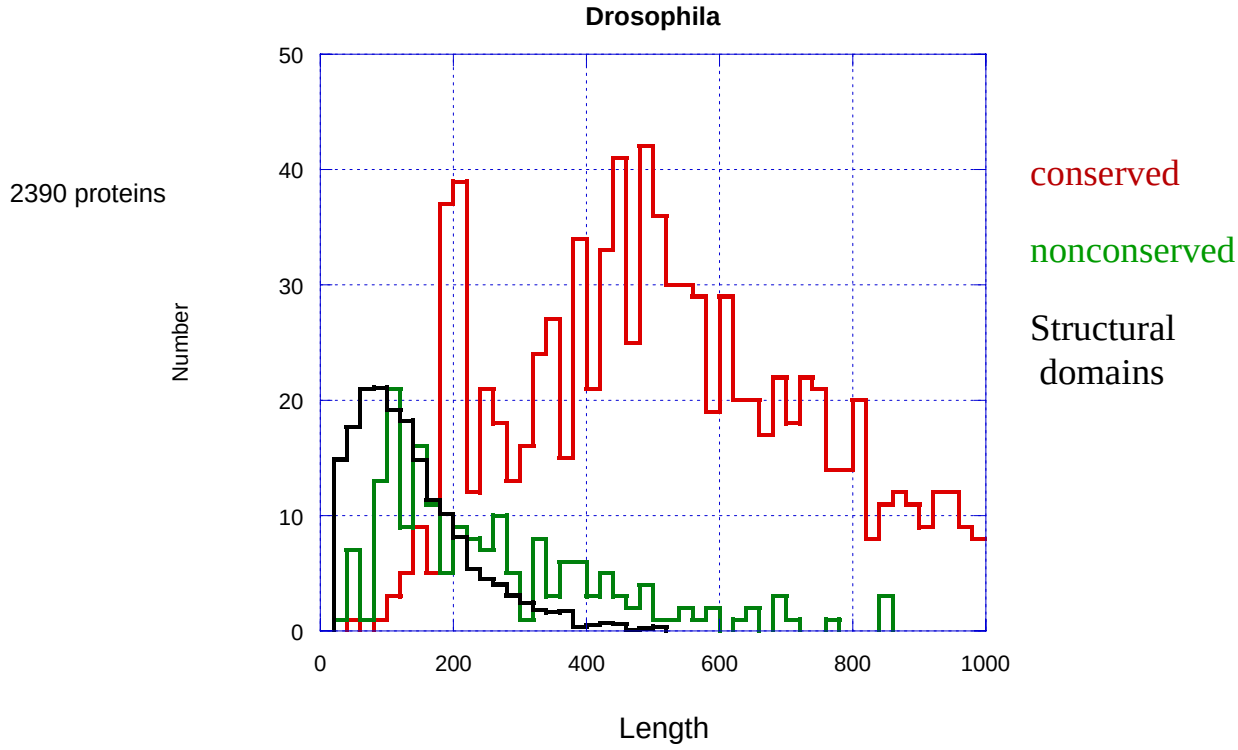

Figure 4
Protein Conservation versus Sequence Length – *Drosophila melanogaster*

Structural Domain Set. In this sense, the poorly conserved proteins from these organisms appear to be as small as proteins can be and still fold into a stable globular structure.

Many biologists implicitly assume that functionally important proteins are more evolutionarily conserved than less vital proteins, and recent work has confirmed this belief [10,11]. Here, we identified another substantial difference between highly conserved and poorly conserved proteins: the less conserved (i.e. less important) proteins are, on average, smaller than more conserved (and more important) proteins. What global evolutionary forces would favor shorter proteins in the absence of other functional constraints? It seems logical to think of these potential forces in terms of minimizing the cost of having extra sequences that do not substantially affect fitness. Such costs might be associated with several distinct processes. One possibility is simply the cost of protein translation and another is the cost of the chaperones that are required to fold longer, particularly multidomain proteins [12]. Although perhaps less likely, yet another cost of longer proteins could be their increased risk of "side effects", i.e. deleterious interactions with other cellular components. For any given protein, the cost differential is likely to be almost negligible, but this difference becomes more significant when one considers the entire set of poorly conserved proteins. In a somewhat similar context, Akashi & Gojobori [13] have shown that highly expressed proteins in the proteomes of *E. coli* and *B. subtilis* have a greater abundance of less energetically costly amino acids than other proteins encoded in these genomes. Another related observation is that of Castillo-Davis and colleagues [14],

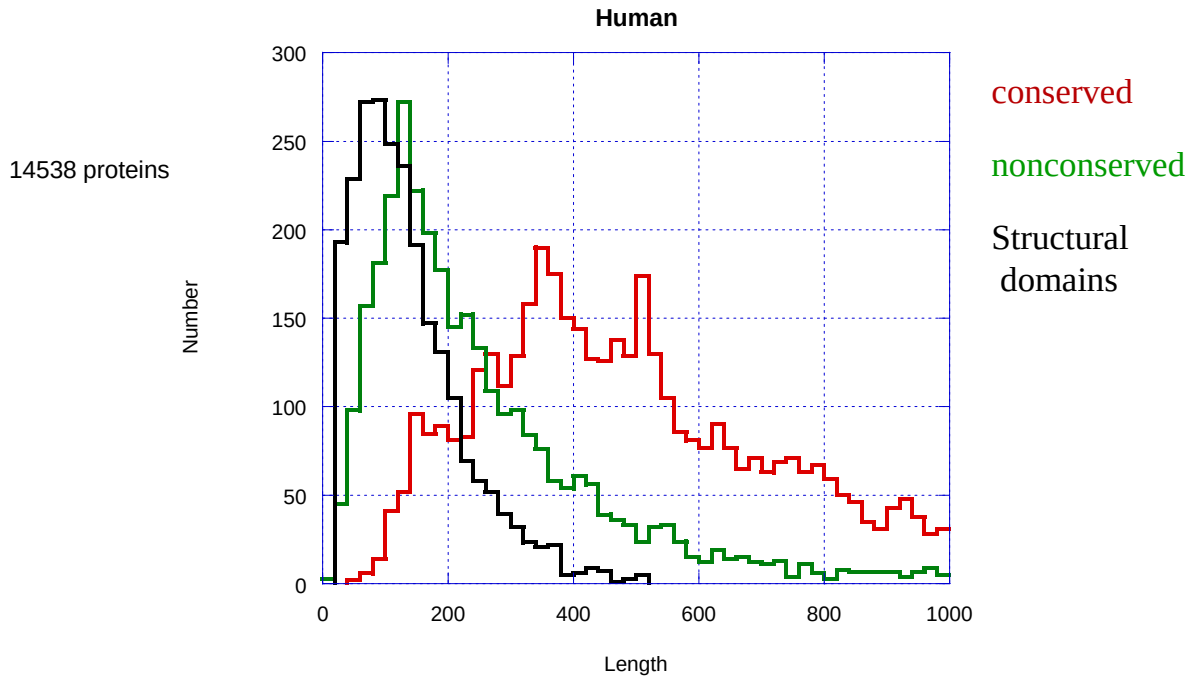

Figure 5
Protein Conservation versus Sequence Length – *Homo sapiens*

who have shown that highly expressed genes have smaller introns on average than other genes presumably due to the cost of transcription and/or splicing.

The action of random genetic drift and selection pressure on genome size (c.f. [15]) could also favor shorter proteins. If deletions are more common than insertions for a given organism, then proteins that can tolerate more mutations (i.e. are evolving under weaker functional constraints) will tend to get smaller over time. Several studies in *E. coli* have indeed shown that, on average, deletions are eight times more frequent than insertions, c.f. [16]. Similarly, analysis of human mutations (A. Kondrashov, personal communication) has shown that deletions are approximately three times more frequent than insertions. It is reasonable to assume that evolutionary forces acting on genome size might be a more important factor favoring smaller proteins for prokaryotic and unicellular eukaryotic genomes because they are primarily composed of protein-coding sequence. This is less obvious for the larger eukaryotic genomes; in particular, the metazoan and plant genomes are primarily composed of noncoding DNA where reductions in protein length would tend to have far less impact on overall genome size.

All of the above constraints would tend to favor shorter proteins but do not seem to explain why the tendency to economize on unnecessary residues increases with greater sequence length, as seen in Figure 7. To have this effect, a constraint must initially have more than a linear increase in intensity with greater sequence length. Given the globular nature of a folded protein, the average number of intramolecular contacts per residue should grow with increasing sequence length (the volume of the globule grows faster than the surface with length increase) and these contacts would constrain the possible residues at any given site within a protein. However, the size of a single globular domain of a protein does not continue to grow with sequence length beyond a certain limit (~150 residues); rather, longer proteins typically have multiple

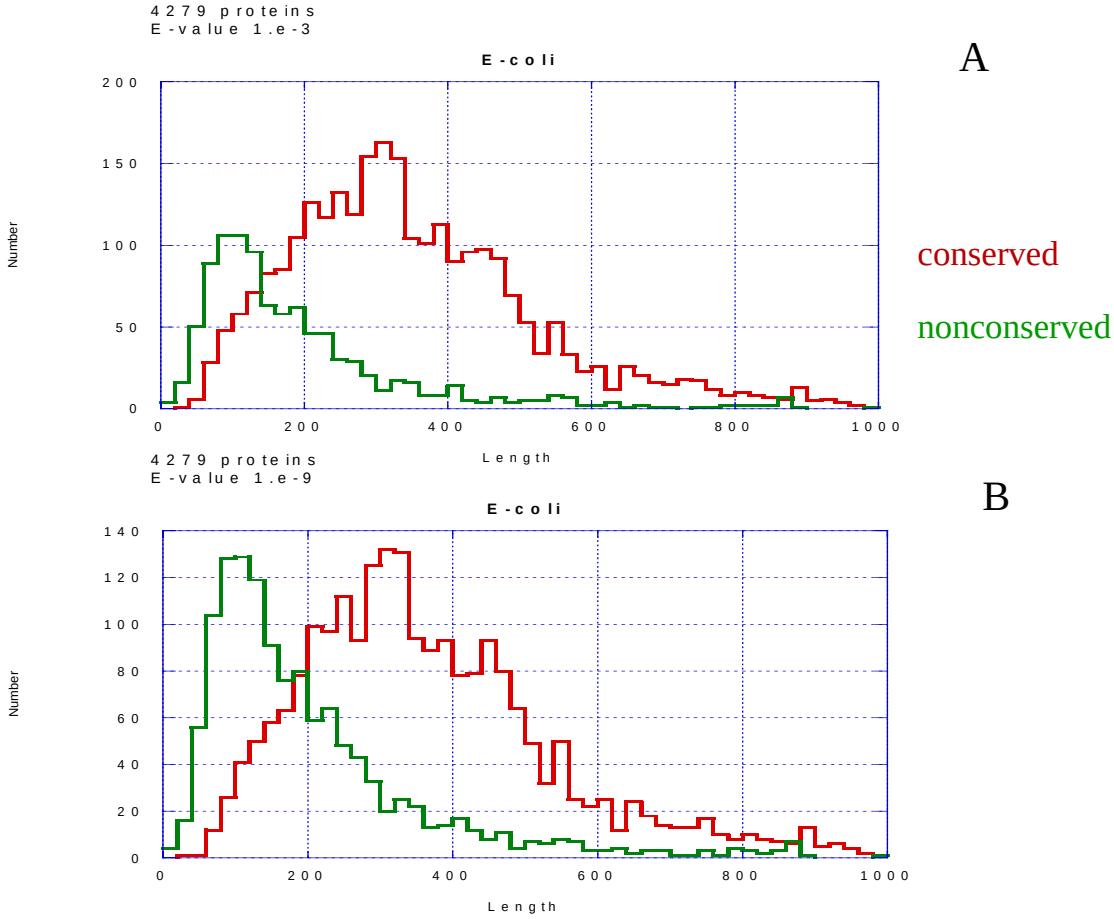

Figure 6
Protein Conservation versus Sequence Length for varying Similarity Thresholds. A. Histograms use similarity threshold of Evalue < 10^{-3} to determine membership in Conserved Set or Nonconserved Set (see Methods). B. Histograms use Evalue < 10^{-9} to determine membership.

globular domains, and thus, the rate of increase in intramolecular contacts for a protein should level off. This is exactly what is seen for the plot of average contact density versus length shown in Figure 7, and the similarity of the contact density plot with that for the fraction of conserved residues is noteworthy. This similarity over a range of sequence lengths is consistent with an evolutionary force minimizing the cost of having extra sequences that do not substantially affect fitness.

The results presented here show that, for all the organisms studied, poorly conserved proteins are, on average, shorter than highly conserved ones. And, in general, there appears to be a significant trend towards shorter proteins in the absence of other, more specific functional constraints. This is compatible with the existence of an evolutionary force acting to minimize the costs associated with sequences that have no functional role. Thus, the size of the poorly conserved proteins seems to tend to minimal domain size, whereas the size of highly conserved proteins varies to a greater extent, reflecting the broad range of functions. It appears that analysis of functionally relatively unimportant proteins allows one to uncover general evolutionary trends that so far remain unnoticed.

Methods

Initial Sequence Sets

The protein sequence sets were derived as follows:

• 4279 *Escherichia coli* protein sequences from NCBI Genomes division, gi NC_002142;

• 2420 *Archaeoglobus fulgidus* protein sequences from NCBI Genomes division gi NC_000917;

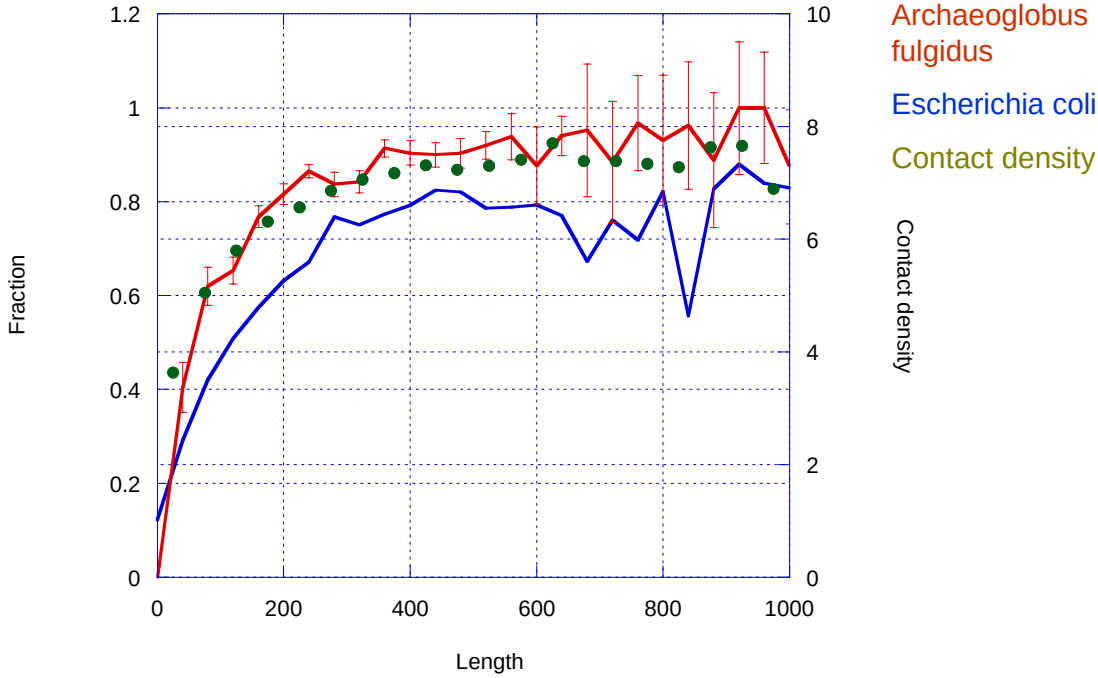

Figure 7
Fraction of Conserved Residues and Amino Acid Contact Density versus Sequence Length – *Escherichia coli* & *Archaeoglobus fulgidus*. **Fraction of conserved residues:** Red line is for *Archaeoglobus fulgidus* and blue line is for *Escherichia coli*. For each bin of 40 residues, the ratio of the total length of non-overlapping cross kingdom alignments divided by the total length of the proteins in that bin is plotted. Proteins that did not have cross kingdom matches contribute zero to the numerator but their lengths are added to the denominator. Error bars shown for *Archaeoglobus fulgidus* were computed as standard error of the mean and were essentially identical for *Escherichia coli* (not shown). **Contact density:** Plotted as green filled circles. For each bin of 50 residues, we plot contact density computed as described in Methods.

• 6305 *Saccharomyces cerevisiae* protein sequences from NCBI Genomes division, gi's NC_001133 – NC_001148, NC_001224, NC_001398;

• 2390 *Drosophila melanogaster* protein sequences extracted from characterized mRNA sequences retrieved from the NCBI Entrez Nucleotides database, requiring a full-length coding sequence and excluding mRNAs generated from high-throughput cDNA projects to minimize partial proteins or proteins generated from ab initio gene predictions;

• 14,538 *Homo sapiens* proteins derived from the NCBI Human RefSeq database, only including proteins encoded by characterized mRNAs and not ab initio gene predictions.

For each organism, protein sequences gained membership to their respective Conserved Sets if they had a BLASTP[17] match of Evalue < 10^{-6} to any sequence in the NCBI nr database from an organism in a different kingdom (i.e. Archaea to Eubacteria, or, Eubacteria to Eukaryota). For the sensitivity tests of this cutoff value, we

repeated the analysis for the *Escherichia coli* proteins using Evalue < 10^{-3} and Evalue < 10^{-9}.

For each organism, protein sequences gained membership to their respective Nonconserved Sets if they had no BLASTP matches of Evalue < 10^{-6} to any sequence in the NCBI nr database (other than within the same organism) or if the only sequence for which they had a BLASTP match was from an organism that was close evolutionarily. Close relatives were defined as follows:

• *Escherichia coli* – Proteobacteria;

• *Archaeoglobus fulgidus* – Euryarcheota;

• *Saccharomyces cerevisiae* – Fungi;

• *Drosophila melanogaster* – Insecta;

• *Homo sapiens* – Mammalia.

Contact Density

A non-redundant set of chain sequences for protein structures from PDB was constructed by single-linkage clustering based on a BLASTP match of Evalue < 10^{-7} or less, as described in Matsuo & Bryant [18]. Contact density was calculated as an average number of contacts per residue (for non-adjacent residues having side chain to side chain distances less than 8 Angstroms).

Acknowledgements

The authors thank L. Wagner and S. Resenchuk for providing data, S. Bryant for input and his suggestion to use the Minimal Structural Domain Set, and R. Doolittle and J. Wilbur for advice.

References

1. Knight RD, Freeland SJ, Landweber LF: **A simple model based on mutation and selection explains trends in codon and amino-acid usage and GC composition within and across genomes.** *Genome Biol* 2001, **2:**RESEARCH0010
2. Galperin MY, Tatusov RL, Koonin EV: **Comparing microbila genomes: how the gene set determines the lifestyle.** *In: Organization of the Prokaryotic Genome (Edited by: Charlebois RL) Washington, DC: ASM Press* 1999
3. Katinka MD, Duprat S, Cornillot E, Metenier G, Thomarat F, Prensier G, Barbe V, Peyretaillade E, Brottier P, Wincker P, Delbac F, El Alaoui H, Peyret P, Saurin W, Gouy M, Weissenbach J, Vivares CP: **Genome sequence and gene compaction of the eukaryote parasite Encephalitozoon cuniculi.** *Nature* 2001, **414:**450-453
4. Singer GA, Hickey DA: **Nucleotide bias causes a genomewide bias in the amino acid composition of proteins.** *Mol Biol Evol* 2000, **17:**1581-1588
5. Wheelan SJ, Marchler-Bauer A, Bryant SH: **Domain size distributions can predict domain boundaries.** *Bioinformatics* 2000, **16:**613-618
6. Madej T, Gibrat JF, Bryant SH: **Threading a database of protein cores.** *Proteins* 1995, **23:**356-369
7. Das S, Yu L, Gaitatzes C, Rogers R, Freeman J, Bienkowska J, Adams RM, Smith TF, Lindelien J: **Biology's new Rosetta stone.** *Nature* 1997, **385:**29-30
8. Skovgaard M, Jensen LJ, Brunak S, Ussery D, Krogh A: **On the total number of genes and their length distribution in complete microbial genomes.** *Trends Genet* 2001, **17:**425-428
9. Lawrence JG, Ochman H: **Amelioration of bacterial genomes: rates of change and exchange.** *J Mol Evol* 1997, **44:**383-397
10. Hirsh AE, Fraser HB: **Protein dispensability and rate of evolution.** *Nature* 2001, **411:**1046-1049
11. Jordan IK, Rogozin IB, Wolf YI, Koonin EV: **Essential genes are more evolutionarily conserved than are nonessential genes in bacteria.** *Genome Res* 2002, **12:**962-968
12. Hartl FU, Hayer-Hartl M: **Molecular chaperones in the cytosol: from nascent chain to folded protein.** *Science* 2002, **295:**1852-1858
13. Akashi H, Gojobori T: **Metabolic efficiency and amino acid composition in the proteomes of Escherichia coli and Bacillus subtilis.** *Proc Natl Acad Sci U S A* 2002, **99:**3695-3700
14. Castillo-Davis CI, Mekhedov SL, Hartl DL, Koonin EV, Kondrashov FA: **Selection for short introns in highly expressed genes.** *Nat Genet* 2002, **31:**415-418
15. Petrov DA: **Evolution of genome size: new approaches to an old problem.** *Trends Genet* 2001, **17:**23-28
16. Halliday JA, Glickman BW: **Mechanisms of spontaneous mutation in DNA repair-proficient Escherichia coli.** *Mutat Res* 1991, **250:**55-71
17. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.** *Nucleic Acids Res* 1997, **25:**3389-3402
18. Matsuo Y, Bryant SH: **Identification of homologous core structures.** *Proteins* 1999, **35:**70-79