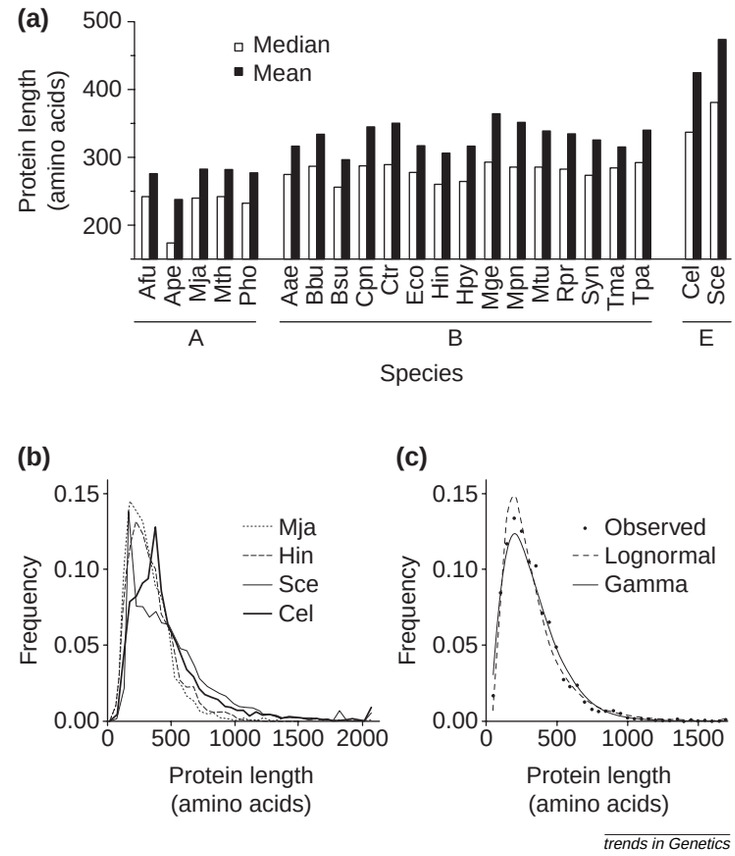
Protein-length distributions for the three domains of life

Protein length varies considerably from dozens to thousands of amino acids. Recent determination of the complete genome sequences of representative organisms from the three domains of life, Archaea, Bacteria and Eukarya, makes it possible for the first time to study the distribution of the length of all proteins encoded in a genome and to compare the distributions across the main lineages of life. Here, I report that the mean protein length is 40–60% greater in eukaryotes than in prokaryotes and discuss the possible biological significance of this phenomenon.

The mean and median lengths of the proteins from 22 species whose genomes have been completely sequenced are presented in Fig. 1a. The means and medians are smallest for archaebacteria and greatest for eukaryotes. For instance, the mean protein lengths (MPLs) for the five archaeal species are between 237 and 282 amino acids, with an average of 270 ±9. This number becomes 330 ±5 for 15 bacteria and 449 ±25 for two eukaryotes. Note that the MPL estimates of eukaryotes are likely to be conservative because of the current limitations of gene-identification tools[1]. Nevertheless, given the fact that a considerable number of proteins are shared among all species[2] and that horizontal gene transfers occur frequently[3–6], the dramatic difference in MPL across the three domains is surprising. The equality of MPL across domains can be rejected statistically ($P < 0.0001$, bootstrap test). This test is possible because the organisms studied within each domain are only phylogenetically distantly related, with the exception of two *Mycoplasma* species and two *Chlamydia* species, and a test without using these species gave the same result. By contrast, within each domain, MPLs are rather similar, despite the distant evolutionary relationships among species. An analysis of variance shows that the variation (mean squares) of MPL among the three domains is about 58 times that found within domains ($F_{[2,19]} = 58$; $P < 10^{-8}$). Similar patterns are observed when the median protein length is considered. Together, these observations suggest that the difference in protein length among domains is not simply because of the independent evolution and accumulation of random mutational changes in the three domains but, rather, that it has biological reasons.

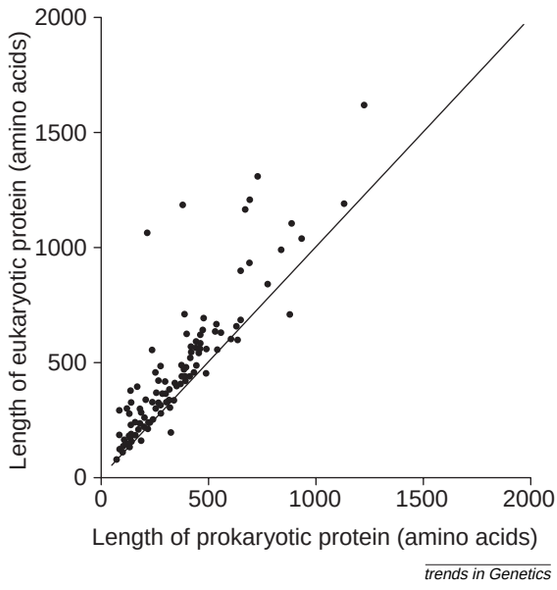
To further characterize the variation in protein length within a species, I computed the protein-length distributions for representative organisms of the three domains: archaebacterium *Methanococcus jannaschii* (Mja), eubacterium *Haemophilus influenzae* (Hin), monocellular eukaryote *Saccharomyces cerevisiae* (Sce), and multicellular eukaryote *Caenorhabditis elegans* (Cel) (Fig. 1b). The distributions for Mja and Hin are quite similar, and they are similar to the distributions for other archaeal and bacterial species examined. But the difference between the prokaryotes and the two eukaryotes is dramatic (Fig. 1b). Apparently, the proportion of big proteins (>500 amino acids) is greater in eukaryotes than in prokaryotes. Interestingly, the distributions for the two eukaryotes also differ substantially, suggesting a distinction between monocellular and multicellular organisms[7], although use of different gene identification tools in analysing the two eukaryotic genomes might also have contributed to the disparity. The observed protein-length distributions can be fitted by the gamma or lognormal

FIGURE 1. Distributions of protein lengths of 22 completely sequenced genomes



(a) Mean and median protein lengths. A, B and E stand for Archaea, Bacteria and Eukarya, respectively. The species names are abbreviated as follows. A: Afu, *Archaeoglobus fulgidus*; Ape, *Aeropyrum pernix*; Mja, *Methanococcus jannaschii*; Mth, *Methanobacterium thermoautotrophicum*; Pho, *Pyrococcus horikoshii*. B: Aae, *Aquifex aeolicus*; Bbu, *Borrelia burgdorferi*; Bsu, *Bacillus subtilis*; Cpn, *Chlamydia pneumoniae*; Ctr, *Chlamydia trachomatis*; Eco, *Escherichia coli*; Hin, *Haemophilus influenzae*; Hpy, *Helicobacter pylori*; Mge, *Mycoplasma genitalium*; Mpn, *Mycoplasma pneumoniae*; Mtu, *Mycobacterium tuberculosis*; Rpr, *Rickettsia prowazekii*; Syn, *Synechocystis PCC6803*; Tma, *Thermotoga maritima*; Tpa, *Treponema pallidum*. E: Cel, *Caenorhabditis elegans*; Sce, *Saccharomyces cerevisiae*. The annotated genome sequences were downloaded from GenBank for all species except Tma, Sce and Cel, which were downloaded from http://www.tigr.org/tdb/mdb/tmdb/tmdb.html, http://www.mips.biochem.mpg.de and http://genome.wustl.edu/gsc/C_elegans/, respectively. (b) Protein-length distributions for Mja, Hin, Sce and Cel. (c) Fitting the protein length variation of Hin by the gamma and lognormal distributions. The shape parameter (α) of the gamma distribution is estimated by m^2/var, where m and var are the mean and variance of the protein-length distribution, respectively. The α estimates for the 22 species are: Afu, 2.27; Ape, 1.94; Mth, 2.09; Mja, 1.98; Pho,1.91; Aae, 2.85; Bbu, 2.19; Bsu, 1.32; Cpn, 2.07; Ctr, 2.04; Eco, 2.33; Hin, 2.29; Hpy, 1.76; Mge, 2.02; Mpn, 1.97; Mtu, 1.66; Rpr, 2.09; Syn, 1.63; Tma, 2.57; Tpa, 2.32; Cel, 1.23; Sce, 1.56. The lower the α value, the greater the variation of the protein length.

FIGURE 2. Comparison of protein lengths



Comparison of the protein lengths of 110 clusters of orthologous genes (COGs) of seven prokaryotes and the yeast. The species used are *Methanococcus jannaschii*, *Escherichia coli*, *Haemophilus influenzae*, *Helicobacter pylori*, *Mycoplasma genitalium*, *Mycoplasma pneumoniae*, *Synechocystis PCC6803* and *Saccharomyces cerevisiae*.

distributions[8] (Fig. 1c) and they fit the data much better than the normal or uniform distributions (data not shown), probably because, for all the genomes examined, the distributions have long tails at the right-hand side. In the case of gamma distribution, the protein length variation can be measured by a shape parameter, α. The lower the α value, the greater the variation. The estimates of α for the 22 species examined are in the range of 1.23 to 2.85, with Cel having the smallest α, indicating that the protein–length variation of Cel is greatest among all species examined. It should be pointed out that fitting observed variations by theoretical distributions is only approximate, as both the lognormal and gamma distributions can be rejected statistically by the goodness-of-fit test. This is not surprising because protein length is affected by numerous factors and there is no particular reason that the empirical distribution should follow any theoretical one exactly.

To examine the evolutionary mechanisms that are responsible for the difference in protein length among domains, I examined whether orthologous proteins are longer in eukaryotes than in prokaryotes. For this purpose, a database of clusters of orthologous groups (COGs)[2] was used. The database compiled putatively orthologous genes in eight species, including archaebacterium Mja, bacteria *Escherichia coli* (Eco), Hin, *Helicobacter pylori* (Hpy), Mge, Mpn and *Synechocystis PCC6803* (Syn), and eukaryote Sce. Although it is extremely difficult to prove gene orthology in such distantly related organisms, members of a COG are obviously homologous if not orthologous, which should serve this purpose as well. There are 110 COGs that each has representative genes from all eight species and the protein length for Sce and the average protein length for the seven prokaryotes were compared for these COGs (Fig. 2). In Sce, 96 COGs show larger protein lengths than in prokaryotes, whereas only the remaining 14 show the opposite. A sign test demonstrates that the length difference between the potentially orthologous proteins of Sce and the prokaryotes is highly significant ($P < 10^{-6}$). The average protein length of these 110 COGs is 359 amino acids for the prokaryotes and 459 for Sce. This difference (100 amino acids) is smaller than the difference (150 amino acids) when all proteins of these species are compared. These results suggest that two factors have contributed to the protein-length difference between eukaryotes and prokaryotes. First, orthologous proteins are longer in eukaryotes than in prokaryotes, and second, eukaryote- and prokaryote-specific proteins are of unequal length on average, with the former being longer than the latter.

It is generally believed that prokaryotic genomes are highly compacted and that only important genes are retained. Bigger proteins might, on average, be less important than smaller ones and therefore are not allowed in prokaryotes, but permissible in eukaryotes. Functional importance of 2971 yeast genes have been tested in gene-knockout experiments, among which, 717 knockouts are lethal and 2254 are viable (http://www.proteome.com). The average protein length of the lethal group is 564 ± 16 amino acids, whereas that of the viable group is 532 ± 8. As the knockout-lethal genes are functionally more important than knockout-viable ones, the above comparison shows that important proteins are slightly larger than less important ones ($P = 0.08$, two-tail Z test), contrary to the hypothesis that bigger proteins are relatively unimportant. A recent study[9] also revealed positive correlation between the protein length and the expression level in both nematode and *Drosophila*. As highly expressed genes are likely to be more important than less-expressed ones on average, their finding is consistent with our result. Another way of evaluating the importance of a gene is to measure the intensity of purifying selection that acts on the gene during evolution, with more important genes being under stronger selective constraints[10]. Unfortunately, selective constraints cannot be measured accurately for many genes of the species studied here because the divergences among the organisms are too large. Therefore, we examined 363 orthologous genes of the mouse and rat that are compiled in Ref. 11. This analysis showed a weak positive correlation (correlation coefficient $= 0.13$, $P = 0.01$) between the protein length and the intensity of purifying selection, suggesting that longer proteins are relatively more important. Here, selection intensity was measured by $1-r$, where r is the ratio of the nonsynonymous nucleotide substitution rate to the synonymous rate[10]. Thus, three lines of evidence suggest that it is not because of relative unimportance that big proteins are not encoded in compacted prokaryotic genomes.

The number of proteins (genes) in higher organisms such as humans is only ~20 times greater than that in *E. coli*. This moderate expansion of the gene inventory in higher organisms might not be sufficient for the tremendous increase in the complexity of the structure and physiology of higher organisms. It is likely that novel gene interactions have contributed greatly to the evolution of higher organisms, and increases in protein length by the addition of functional motifs can be an important evolutionary strategy for achieving sophisticated gene regulation networks in eukaryotes. In summary, although the biological meaning and evolutionary mechanism of the difference in protein sizes among the three domains remain elusive at present, they can be studied with biochemical and computational techniques that are currently available, and I believe it will not be long before we solve these mysteries.

Jianzhi Zhang
jzhang@niaid.nih.gov

Laboratory of Host Defenses, National Institute of Allergy and Infectious Diseases, National Institutes of Health, Building 10, Room 11N104, 9000 Rockville Pike, Bethesda, MD 20892, USA; and Institute of Molecular Evolutionary Genetics, Pennsylvania State University, PA, USA.

Acknowledgements

I thank M. Nei, I. Rogozin, A. Rooney, and H. Rosenberg for discussions.
This work was partly supported by the NIH and NSF research grants to M. Nei.

References

1 Das, S. *et al.* (1997) Biology's new Rosetta stone. *Nature* 385, 29–30
2 Tatusov, R.L. *et al.* (1997) A genomic perspective on protein families. *Science* 278, 631–637
3 Brown, J.R. and Doolittle, W.F. (1997) *Archaea* and the prokaryote-to-eukaryote transition. *Microbiol. Mol. Biol. Rev.* 61, 456–502
4 Koonin, E.V. *et al.* (1997) Comparison of archaeal and bacterial genomes: computer analysis of protein sequences predicts novel functions and suggests a chimeric origin for the Archaea. *Mol. Microbiol.* 25, 619–637
5 Brown, J.R. *et al.* (1998) A bacterial antibiotic resistance gene with eukaryotic origins. *Curr. Biol.* 8, 365–367
6 Nelson, K.E. *et al.* (1999) Evidence for lateral gene transfer between Archaea and Bacteria from genome sequence of *Thermotoga maritima*. *Nature* 399, 323–329
7 Chervitz, S.A. *et al.* (1998) Comparison of the complete protein sets of worm and yeast: orthology and divergence. *Science* 282, 2022–2028
8 Johnson, N.L. and Kotz, S. (1970) *Distributions in Statistics: Continuous Univariate Distributions*, Houghton Mifflin
9 Duret, L. and Mouchiroud, D. (1999) Expression pattern and, surprisingly, gene length shape codon usage in *Caenorhabditis*, *Drosophila*, and *Arabidopsis*. *Proc. Natl. Acad. Sci. U. S. A.* 96, 4482–4487
10 Nei, M. (1987) *Molecular Evolutionary Genetics*, Columbia University Press
11 Wolfe, K.H. and Sharp, P.M. (1993) Mammalian gene evolution: nucleotide sequence divergence between mouse and rat. *J. Mol. Evol.* 37, 441–456

Regulation of adjacent yeast genes

Semyon Kruglyak
kruglyak@hto.usc.edu

Haixu Tang
tanghx@hto.usc.edu

Department of Mathematics, University of Southern California, Los Angeles, CA, USA.

In the genomes of prokaryotes many cases are known where a single regulatory system controls two or more functionally related genes[1]. Although important differences exist between the regulatory systems of prokaryotes and eukaryotes, it has been suggested that multi-gene regulation also exists in eukaryotic genomes. We explore this notion through the analysis of gene-expression data because observed changes in gene-expression levels quantify the effect of the regulatory mechanisms. The yeast *Saccharomyces cerevisiae* is our model organism because the whole genome has been sequenced, all open reading frames have been identified, and much expression data is available.

In our discussion of the regulatory system, we are referring to the transcription-factor-binding sites that interact with various proteins to regulate gene expression at the transcription level. It is widely accepted that yeast regulatory systems are typically located within several hundred base pairs (bp) upstream of the genes they control[2–4]. Typical searches encompass 800 bp upstream of the start codon[5]. Zhang and Smith have found many functionally related genes that are located near one another in the yeast genome[6] and Cho *et al.*[7] show the existence of adjacent genes the expression of which is initiated in the same phase of the cell cycle. A disproportionate fraction of these genes are transcribed on opposite strands, away from each other[7]. This would suggest that a regulatory system that was located between the two genes could control the expression of both.

This is interesting in itself, and it might lead to increased understanding of the regulatory system. Finding transcription-factor-binding sites[8] and deciphering their interactions[9] are important and difficult problems. The case of a single regulatory system controlling a gene pair is easier to analyse than the general case because the search for transcription-factor-binding sites can be narrowed to the region between the two genes. There are many examples where this region is less than 400 bp in length (see Table 1). In addition, the fact that two genes must be controlled could impose some symmetry on the regulatory system. For example, symmetry might be reflected in the interaction of transcription factors or by the presence of palindromic binding sites. The known binding sites in the region of interest can be identified by resources such as the TRANSFAC database[10].

In this work, we present additional evidence for the existence of multi-gene regulation in yeast, and we give a list of candidate gene pairs that are likely to be controlled in this manner. Our conclusions are based on expression data from cell cycle[7], diauxic shift[11] and sporulation experiments[12]. In these experiments, microarray technology has been used to measure the expression level of every yeast gene at a series of time points, and in each experiment, the time horizon spans one of the biological processes mentioned above. We compare the expression patterns of adjacent genes in these data sets.

The key idea is that two genes that are controlled by a single regulatory system should have similar expression patterns in any data set. We used the correlation coefficient as a measure of similarity to show that the expression patterns of adjacent genes are more often highly correlated than the expression patterns of randomly selected gene pairs. Because the correlation coefficient is highly sensitive to experimental variation in the data, we filtered the data sets to include only genes whose expression values undergo substantial changes during the time course of the experiment. Such genes are informative because their expression patterns have a high signal to noise ratio.

TABLE 1. Candidate pairs for control by a single regulatory system[a]

ORF	Sporulation	Diauxic shift	Cell cycle	Direction	Distance
YAR007C YAR008W	0.82	0.83	0.96	← →	345
YBR052C YBR053C	0.73	0.83	0.90	← ←	322
YDR229W YDR230W	0.76	0.83	0.77	→ →	86
YIL020C YIL019W	0.87	0.85	0.87	← →	282
YJL190C YJL189W	0.73	0.97	0.91	← →	630
YKR024C YKR025W	0.94	0.63	0.86	← →	397
YLL062C YLL061W	0.67	0.81	0.93	← →	342
YNL294C YNL293W	0.80	0.72	0.73	← →	379
YNL263C YNL262W	0.78	0.78	0.67	← →	371
YNL037C YNL036W	0.84	0.81	0.73	← →	811

[a]Candidate gene pairs that had correlated expression patterns in all three data sets. Direction refers to the direction of transcription of each gene in the pair, and distance is the number of base pairs between the genes.

 PII: S0168-9525(99)01941-1