Proc. Natl. Acad. Sci. USA
Vol. 83, pp. 1198–1202, March 1986
Biochemistry

Proteins of *Escherichia coli* come in sizes that are multiples of 14 kDa: Domain concepts and evolutionary implications

(distribution of protein sizes/HeLa cell proteins/distribution of gene lengths/protein structure/chromatin structure)

MICHAEL A. SAVAGEAU

Department of Microbiology and Immunology, The University of Michigan, Ann Arbor, MI 48109

Communicated by J. L. Oncley, October 10, 1985

ABSTRACT **Initial attempts to correlate the distribution of gene density (number of gene loci per unit length on the linkage map) with the distribution of lengths of coding sequences have led to the observation that 46% of approximately 1000 sampled proteins in *Escherichia coli* have molecular masses of *n* × 14,000 ± 2500 daltons (*n* = 1, 2, ...). This clustering around multiples of 14,000 contrasts with the 36% one would expect in these ranges if the sizes were uniformly distributed. The entire distribution is well fit by a sum of normal or lognormal distributions located at multiples of 14,000, which suggests that the percentage of *E. coli* proteins governed by the underlying sizing mechanism is much greater than 50%. Clustering of protein molecular sizes around multiples of a unit size also is suggested by the distribution of well-characterized HeLa cell proteins. The distribution of gene lengths for *E. coli* suggests regular clustering, which implies that the clustering of protein molecular masses is not an artifact of the molecular mass measurement by gel electrophoresis. These observations suggest the existence of a fundamental structural unit. The rather uniform size of this structural unit (without any apparent sequence homology) suggests that a general principle such as geometrical or physical optimization at the DNA or protein level is responsible. This suggestion is discussed in relation to experimental evidence for the domain structure of proteins and to existing hypotheses that attempt to account for these domains. Microevolution would appear to be accommodated by incremental changes within this fundamental unit, whereas macroevolution would appear to involve "quantum" changes to the next stable size of protein.**

Analysis of gene density on the genetic map of *Escherichia coli* has shown that the frequency distribution is lognormal (1)—that is, a skewed distribution with low densities occurring frequently and high densities occurring infrequently. It was concluded that such a skewed distribution was unlikely to be an artifact due to the small sample of genes that have been mapped. The lognormal character of the frequency distribution was found for both the 1976 map (2) and the 1983 map (3), which included many more loci. More important, it was argued that the skewed distribution is fundamental in the sense that the distribution of protein sizes, and hence their coding sequences, appears to be lognormal. This is suggested by visual inspection of typical O'Farrell gels (4). Since the migration distance in the $NaDodSO_4$/PAGE dimension is approximately proportional to the logarithm of protein molecular mass, the nearly normal distribution of protein spots suggests a logarithmic distribution of molecular masses.

Size Distribution of *E. coli* Proteins

The new protein catalog for *E. coli* (5) has made it possible to refine the protein size distribution. Approximately 1000 different proteins (polypeptide chains) have been isolated on O'Farrell gels and associated with a region of the genome containing their coding sequence. When the frequency distribution of molecular masses for these proteins is plotted, one finds the expected skewed distribution (1).

Further analysis with more refined class sizes for molecular mass shows that although the "envelope" of this distribution is nearly lognormal, there is a pronounced "fine structure" (Fig. 1). A class size of 2 kDa was chosen for this distribution because it smooths out random variations on the scale of experimental error while retaining the fine structure expressed on a scale larger than this error. A periodicity of about 14 kDa is seen. Forty-six percent of the proteins have molecular masses of *n* × 14 ± 2.5 kDa (*n* = 1, 2, ...). This clustering around multiples of 14 kDa contrasts with the 36% one would expect in these ranges if the sizes were uniformly distributed. This period of 14 kDa is evident for four cycles and then disappears as the data at higher molecular masses become sparse; the clustering per cycle is 51% at *n* = 1, 44% at *n* = 2, 42% at *n* = 3, 52% at *n* = 4, and 37% at *n* > 4. The same periodic pattern is evident when class sizes of 1, 2, ..., 7 kDa are used, but disappears, as expected, when class sizes greater than or equal to the period of 14 kDa are used. The period also is unaffected by the phase with which the smaller class sizes are laid out; i.e., whether the first interval is started at 0, 1, 2, ... kDa. Minima in these distributions occur reproducibly at 23, 37, 51, and 65 kDa. These results are summarized in Table 1. Thus, the 14-kDa period is not the result of a fortuitous choice of class size or phase.

If the four peaks are represented by the sum of four normal distributions, and one estimates the mean, standard deviation, and contribution (or amplitude) of each to the net distribution, then one obtains the results given in Table 2. These are the estimated parameter values, with the corresponding standard deviations in the estimates, that produce the best fit to the empirical distribution in the sense of least-squared error. The estimation was done by computer using the Gauss–Newton method. The estimated means for the four peaks reinforce the notion of a fundamental 14-kDa period. If one assumes a sum of four lognormal distributions, rather than normal distributions, one obtains essentially identical values for the means. If the frequencies are weighted according to their molecular mass, again one obtains essentially identical estimates for the means. Fig. 2 shows the empirical data for molecular masses from 0–66 kDa and the estimated distribution based upon the sum of four normal distributions. For emphasis, the sparse and more erratic data at higher molecular mass have been omitted from this figure.

Size Distribution of Animal Cell Proteins

The size distribution of total HeLa cell and other animal cell protein (6, 7) is skewed and generally similar to that of *E. coli.*

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

Abbreviation: bp, base pair(s).

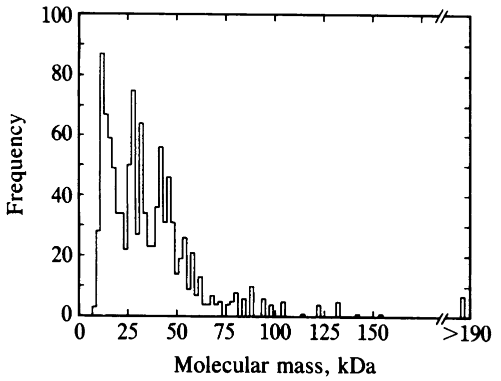

FIG. 1. Multimodal frequency distribution of protein sizes for *E. coli*. Molecular masses were estimated from migration distance in two-dimensional gels (5) by piece-wise logarithmic interpolation between (and extrapolation from extreme) reference markers, since migration distance is not strictly related to the logarithm of molecular mass. The reference markers used were 10, 12, 15, 20, 25, 30, 35, 40, 45, 50, 60, 80, 100, and 160 kDa, corresponding to standard migration distances of 10, 20, 32.3, 48, 60, 70, 78, 85, 91, 97.5, 103.5, 112, 116.5, and 122.5 mm (see figure 1 of ref. 5). The molecular masses calculated in this fashion for a sample of 56 proteins whose genes have been sequenced are 99.4% correlated with their gene lengths (see text). The same form of distribution is obtained whether or not the proteins with an asterisk in table 5 of ref. 5 are included. See text for discussion.

Although the periodicity of 14 kDa in the *E. coli* distribution can be detected in one of the earlier studies (6), it was not noted by the authors. Since the number of distinct proteins (and therefore coding sequences) represented by a given molecular mass sample was not determined in this earlier study, one might have attributed the fine structure to differential expression of coding sequences with a uniform size distribution. The HeLa cell distribution in this same study (6) has a less pronounced fine structure. A better test of fine structure in the distribution of HeLa cell proteins could be obtained by analysis of data from two-dimensional gel electrophoresis. Although the published catalog of HeLa cell proteins is not as extensive as that for *E. coli*, 99 major proteins have been well characterized (8). The frequency distribution of molecular masses for 88 of these proteins is shown in Fig. 3. The 11 largest proteins, which constitute a very sparse set of data, have been omitted. The distribution suggests a clustering of proteins at 19, 32, 46, 57, and 68 kDa. All relevant class sizes and phase relationships have been examined, and minima in the distributions occur reproducibly at 24, 40, 61, and 79 kDa. These results suggest a clustering of protein molecular masses not unlike that seen for *E. coli* proteins, although perhaps shifted toward somewhat higher molecular masses. However, one should not place too much emphasis upon the numerical values of the peaks given above, since the data are limited and no statistically significant estimates of the means and standard deviations for the peaks could be obtained.

Size Distribution of *E. coli* Genes

Although care was taken to ensure that the periodicities seen in Figs. 1–3 are not due to the manner of plotting the data, one might ask whether the two-dimensional gel methods themselves could introduce such artifacts. There is no evidence to suggest that they do lead to clustering of protein spots, but a more direct and independent test would be provided by the lengths of coding sequences in the genome.

The lengths of all *E. coli* coding sequences currently in GenBank were determined and their distribution was plotted (Fig. 4). Leader and signal sequences and precursor forms were omitted as well as the largest 14 genes that constitute too sparse a data set, leaving a total of 146 genes. The distribution suggests a clustering of lengths at about 400, 800, 1200, and

Table 1. Minima in the frequency distributions of protein molecular masses for *E. coli*

PS*	Positions of minima[†] (kDa), with associated ratios[‡]							
				Class size: 1 kDa				
0	23.5	30.5	36.5	45.5	50.5	55.5	59.0	65.5
	217/7	50/7	249/11	71/9	13/4	24/8[§]	18/7[§]	18/4[§]
				Class size: 2 kDa				
0	23.0		37.0	45.0	53.0		59.0	65.0
	248/17		222/19	44/23	58/9		8/6	9/3[§]
1	24.0	30.0	37.0		50.0	56.0	60.0	65.0
	336/18	58/22	196/17		32/11	20/7	8/6	12/3
				Class size: 3 kDa				
0	22.5		37.5		52.5			
	198/21		153/17		42/10			
1	22.5		34.5	43.5	49.5		58.5	64.5
	153/18		87/21	42/20	24/11		15/4	12/2
2	24.5		36.5					65.0
	252/21		93/17					24/2
				Class size: 4 kDa				
0	22.0		38.0		50.0			66.0
	104/24		88/22		12/12			8/2
1	23.0		35.0					
	180/22		92/23					
2	24.0		36.0		52.0			66.0
	196/22		96/21		40/12			8/3
3	21.0		37.0					65.0
	104/27		76/17					4/3
				Class size: 5 kDa				
0	22.5		37.5					
	145/22		55/19					
1	23.5		38.5					68.5
	125/25		40/21					10/3
2	24.5		39.5					64.5
	145/26		145/21					20/2
3	20.5		35.5			55.5		
	90/27		50/21			20/13		
4	21.5		36.5		51.5			
	95/23		60/18		35/10			
				Averages				
	22.8	30.3	36.8	44.7	51.1	55.7	59.1	65.5

*Phase shift in kDa.

[†] Minima are recorded when they have an associated ratio ≥1.

[‡] Area of the "valley" from the value of the lowest adjacent peak down to the value of the minimum, divided by the value of the minimum.

[§] Minima with value zero are arbitrarily given a value equal to the minimum of the nearest nonzero neighbor, to avoid infinite ratios.

Table 2. Parameter values for the four normal distributions producing the best fit to the net distribution of protein molecular masses for *E. coli*

Component* (i)	Mean (μ_i)	Standard deviation (σ_i)	Amplitude[†] (A_i)
1	13.90 ± 0.605	3.168 ± 0.601	612.26 ± 113.19
2	28.65 ± 0.937	5.481 ± 1.035	756.30 ± 116.87
3	43.13 ± 1.447	3.374 ± 1.647	348.03 ± 328.42
4	53.85 ± 8.287	6.511 ± 6.615	302.52 ± 338.74

*$F = \Sigma_{i=1}^{4} A_i[(2\pi)^{1/2}\sigma_i]^{-1} \exp[-(M - \mu_i)^2/2\sigma_i^2]$, where M is the molecular mass.

[†] These numbers are twice the actual values because frequencies are counted twice with a class size of two.

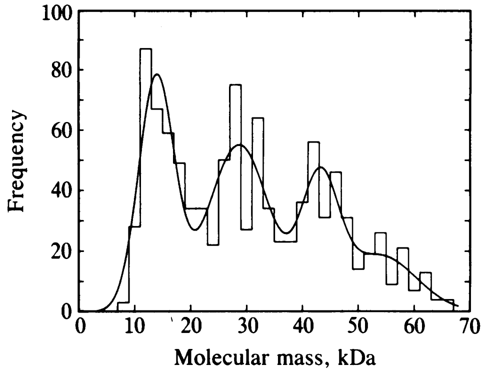

FIG. 2. Multimodal frequency distribution of protein sizes for *E. coli*. The data in Fig. 1 between 0 and 66 kDa are replotted on an expanded scale for emphasis. The smooth curve is the best-fitting distribution composed of the sum of four normal distributions. The estimated values for the parameters of the four distributions are given in Table 2.

1600 base pairs (bp), corresponding approximately to proteins of 14.6, 29.3, 44.0, and 58.6 kDa (see below). All relevant class sizes and phase relationships were examined, and minima in the distributions occur reproducibly at 650, 975, 1475, and 1850 bp. Again, one should not place too much emphasis on the numerical values of the peaks given above, since the data are too limited to obtain statistically significant estimates for their means and standard deviations.

A plot of molecular mass [estimated from migration distance in two-dimensional gels (5) by piece-wise logarithmic interpolation between reference markers] against gene length for a sample of 56 proteins that have been identified in two-dimensional gels and whose genes have been sequenced shows the two measurements to be highly correlated (correlation coefficient = 0.994). The samples include the full range of molecular mass and basic as well as acidic proteins. The slope of the regression line indicates an average molecular mass of 110 ± 1.5 Da per amino acid residue. Thus, the full scale of 70 kDa in Fig. 2 corresponds closely to the full scale of 1900 bp in Fig. 4.

These results on gene lengths reinforce the conclusion based on protein sizes and argue that the 14-kDa periodicity is not an artifact introduced by the electrophoretic methods. However, at this time the best data for making a statistically significant estimate of the preferred protein sizes is still the two-dimensional gel data for *E. coli*, where the data are 10-fold more abundant.

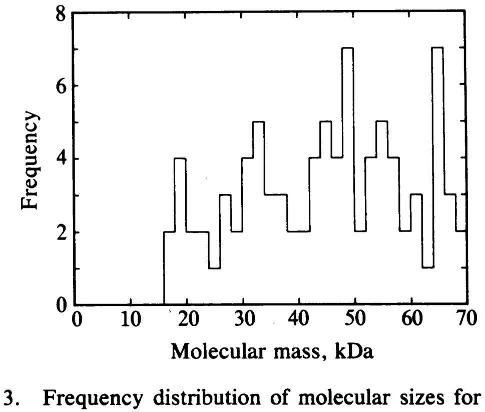

FIG. 3. Frequency distribution of molecular sizes for 88 well-characterized proteins from HeLa cells. Molecular masses are published values estimated from migration distance in two-dimensional gels (8).

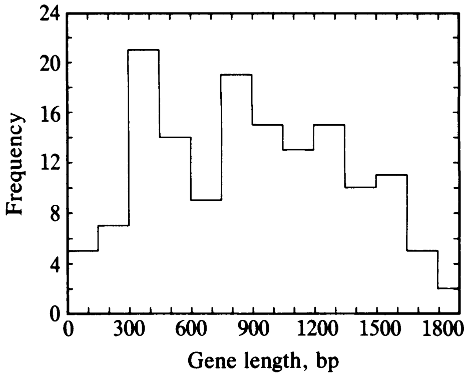

FIG. 4. Frequency distribution of gene lengths for *E. coli*. All but the 14 largest lengths for *E. coli* genes of known sequence are shown. A class size of 150 bp was selected to diminish the random variation associated with the small sample of genes.

A Structural Unit of 14 kDa

The most straightforward interpretation of these results is that a basic structural unit of about 14 kDa exists.* A large fraction of *E. coli* proteins could be composed of such 14-kDa units, perhaps much greater than 50%. Two types of explanations can be posited to account for this observation: accidental fixation or selection. According to the first type, there was a primordial protein of about 14 kDa from which the current proteins of *E. coli* have evolved. Larger proteins undoubtedly have been constructed by duplication and assembly of such units. Proteins smaller than 14 kDa may have evolved from the 14-kDa unit size or independently. In any case, the evolution of proteins by assembling the basic units and possibly such smaller peptides could yield the periodic pattern seen in Figs. 1 and 2. According to this type of explanation, the DNA sequences for these proteins should show considerable homology. However, the fact that there are no data to indicate a corresponding degree of repeated sequence homology or families of multiple homologous sequences in the *E. coli* genome suggests that the sequences of the 14-kDa units are not highly conserved. In the absence of other factors, any periodic pattern should disappear with time because of the leveling influence of random insertions and deletions. Alternatively, according to the second type of explanation, the clustering of sizes could be the result of selective pressures operating at the DNA or protein levels. Some mechanisms that might account for such selection are discussed in the following section. In the case of either type of explanation, the question remains: What determines the unit size and its dispersion?

The rather uniform size of this basic unit in the apparent absence of sequence uniformity suggests that a general principle such as geometrical (sizing, positioning) or physical (stabilizing) optimization at the DNA or protein level is

*It is interesting that such a structural unit was originally suggested on the basis of a small sample of proteins by several investigators in the period between 1930 and 1940. The hypothesis that globular proteins have molecular weights that are multiples of a basic unit was put forward by Svedberg in 1929 and later refined to the formula (9) molecular weight = $n \times 17{,}600$, where $n = 1, 2, 4, 8, 16, 48, 96, (168), 192, 384, 576$. This molecular weight sequence also was suggested in proposals by Wrinch (10), whose "cyclol hypothesis" predicted stable three-dimensional structures with 144 and 288 amino acid residues, and Bergmann and Niemann (11), whose hypothesis of periodicity of amino acids in peptide chains led them to suggest stable protein structures of $n \times 288$ amino acid residues. These proposals were refuted strongly by a number of workers, and the status of this work was well-reviewed in a number of classical texts of this period (12–14).

responsible. Existing hypotheses for domain structure at the DNA or protein level provide a number of specific suggestions.

Domain Concepts

Folding Domains. Multiple domains of relatively compact structure in globular proteins have been noted for some time (15–18). Wetlaufer (16, 17) has argued that such domains permit simultaneous folding of independent regions and thus facilitate the folding of native proteins on a biologically acceptable time scale. The small sample of proteins from which these observations have been drawn has not suggested a uniform size for such domains. Richardson (18) noted that, although the commonest domain size is between 100 and 200 residues, there appears to be no definite upper limit. The range observed covers an order of magnitude (40–400) and varies according to the major structural category of the proteins being examined. Wetlaufer (16) observed a range from about 20–40 to 150 amino acid residues and speculated that this size range may result from a relationship between "nucleation time" and "growth time" for the domains. Karplus and Weaver (19), however, argued against this "random-search nucleation and chain propagation" mechanism for protein folding, at least for domains in the range of concern here. The alternative they suggested is a mechanism based on "diffusion and collision," but it is not clear that this mechanism would predict an optimum size of 14 kDa for these domains.

More subtle structural modules within proteins that lack domain or supersecondary structure, as defined above, have been reported by Go (20, 21). The sizes of these modules (20–40 amino acids) are at the lower limit of the putative folding domains suggested by Wetlaufer (17).

Immunoglobulin Domains. Another domain concept has developed in immunology following the structural determination of immunoglobulin proteins (22–24). The domains in this case correspond to exons in the genome (25, 26). Other, non-immunoglobulin components of the immune system have been shown to possess similar domains, and their sequence relatedness has suggested an evolutionary relationship among these molecules (27–31). This family also may include proteins associated in a general way with the function of cell recognition (28). These appear to have an internal disulfide-bonded loop approximately the size (10 kDa) of an immunoglobulin domain and to be members of a family related by sequence (28–33). The sequence relatedness has suggested a specific functional rationale and evolutionary relationship for these domains (28), which may be difficult to justify in the general case of *E. coli* proteins.

Functional Domains. It is well-known that most active centers are located in clefts on the surface of protein molecules and at the interface between domains. A particularly convincing explanation has been given for the location of centers between domains in α/β proteins (34). Recent studies of the reactivity of antipeptide antibodies have suggested that mobility of one domain relative to another may be an essential part of protein–protein recognition in general (35, 36). There is no evidence that these functional requirements would lead to the 14-kDa unit described here.

Exon Domains. There may be alternative explanations based upon the mechanisms of gene rearrangement or the physical chemistry of chromatin. A common hypothesis is that genes evolve by the duplication and shuffling of relatively conserved exons (37, 38), which further suggests a mosaic arrangement of functional domains for proteins (25, 39–41). Could this hypothesis alone account for the 14-kDa unit? I don't believe so (even if we were to grant the existence of such a shuffling mechanism in *E. coli*). One would have to postulate some additional mechanism to account for the single, strongly conserved size of 14 kDa. The sizes of exons often correspond to polypeptides that are much smaller than 14 kDa (21, 42, 43), and the polypeptides do not always correspond to functional domains in proteins (41).

On the other hand, since many exons have sizes that tend to cluster around 120–150 bp (polypeptide mass 4.4–5.5 kDa) (42, 43), the 14-kDa unit might be composed of two or three protein modules of the type defined by Go (21). Composites of two or three modules per subunit might be expected, since 75% of all proteins appear to be composed of either two or four subunits (7). Each subunit of such a multimeric protein must have at least two functional domains—one for binding to other subunits and one for expressing a function associated with the native protein. Larger polypeptides, which are composed of multiples of the 14-kDa unit, would presumably have arisen by fusion of such 14-kDa units.

Because many exons have sizes that tend to cluster around 120–150 bp (42, 43), one might expect to find this reflected in the protein size distribution, particularly at lower molecular mass. Fig. 2 shows no evidence of a 6-kDa periodicity within the first peak at 14 kDa. There is an apparent splitting of the peaks at 28, 42, and perhaps 56 kDa, but this depends upon heavily weighting single experimental values. The significance of this apparent splitting is difficult to assess because the empirical data themselves are not always accurate to within 1–2 kDa and because a shifting of the class intervals by 1 kDa in phase can abolish the splitting seen in Fig. 2. It is possible that proteins from eukaryotic cells, in which splicing mechanisms are more prominent, might display a periodicity of 6–8 kDa. However, the limited data for HeLa cells in Fig. 3 and in plots with alternative phases do not reproducibly suggest a periodicity in this range.

Chromatin Domains. At the level of chromatin structure, a repeat unit of 380-bp length might have selected for proteins with the corresponding domain structure. However, the physical basis of such putative 380-bp units, and whether such a mechanism would be more fundamental than those acting directly on protein structure, would have to be established.

In the case of eukaryotic chromatin, there is good evidence for a dinucleosome structure (44) consisting of 1 and 3/4 turns of DNA (146 bp) plus linker DNA (22–100 bp, depending upon cell type) per nucleosome (45). Thus, the dinucleosome unit totals 336–492 bp of DNA. Although alternative models of chromatin structure based on the dinucleosome unit exist (45–47), recent evidence is consistent only with the "crossed-linker" model (45). For the structure of this chromatin unit to be reflected in the sizes of mature proteins, in the most straightforward interpretation, would require exons of a commensurate size (and perhaps suitable phasing). However, as noted above, exon sizes have been reported to cluster around 120–150 bp (42, 43). There are some obvious implications that must be tested before these observations can be reconciled.

At present, the evidence for such structures in the nucleoids of *E. coli* is less certain but still suggestive. Fragments of DNA resistant to DNase I digestion have sizes of about 120 bp, not unlike that associated with eukaryotic nucleosomes (48). Electron micrographs have suggested a rather fragile, chromatin-like structure with condensation of DNA in blocks of about 250 bp (49).

Interesting experiments regarding the mechanisms of recombination/transcription/translation/protein stability, as a function of gene length, are suggested by these units of DNA and protein organization.

Evolutionary Implications

A strongly conserved structural unit for proteins would have important implications for evolution. Amino acid substitutions and small insertions/deletions could lead to incremental changes within existing proteins and thus microevolution. This would contribute to the dispersion in the peaks of Fig. 2. However, one would not expect larger changes (macroevolution) to occur by the accumulation of small incremental changes in existing proteins. Beyond certain sizes, the proteins would be unstable or otherwise strongly selected against. Instead, macroevolutionary "leaps" would occur by "quantum" changes to the next stable size of protein. Entirely new functions, characteristic of a complete polypeptide domain, could be added to a protein in this fashion, and in the case of regulatory proteins, the ability to activate an entire developmental program might be acquired.

These general ideas have often been suggested in connection with gene duplication and shuffling; what I am suggesting here is that perhaps another mechanism, having to do with sizing or stability and leading to the 14-kDa unit, might have a profound selective influence on such processes.

Conclusion

The distribution of *E. coli* proteins shows a distinct periodic pattern with sizes clustering around multiples of about 14 kDa. A similar pattern is suggested for HeLa cell proteins, although the data presented are less extensive. The periodic pattern in the size distribution of genes from *E. coli* shows that the periodicity in the size distribution of proteins is not simply an artifact of molecular mass measurement by gel electrophoresis.

There is experimental evidence for domains of various sizes. Although there are hypotheses that account for many of these, none provides a completely satisfactory explanation for the 14-kDa protein unit reported in this paper. The strongly conserved size, together with the apparent lack of sequence homology, suggests that the 14-kDa units exist for some fundamental geometrical or physical reason and then serve as "frameworks" for the construction of essentially all functional proteins. In some cases, the framework itself also might possess functional properties; in other cases, the functions might be "added" to the framework at junctions between 14-kDa units of the framework or at the ends of such units. Whatever the physical basis for this strongly conserved 14-kDa structural unit, it has important implications for our understanding of molecular evolution, particularly the relationship between micro- and macroevolutionary events.

I thank Drs. Rod Nairn, Martha Ludwig, C. W. Birky, and John Langmore for helpful discussions and suggestions, Dr. E. O. Voit for assistance in the computer analysis of the frequency distributions, and Dr. J. L. Oncley and an anonymous reviewer for making helpful suggestions in the revision of the manuscript. This work was supported in part by a grant from the National Institutes of Health (GM30054).

1. Jurka, J. & Savageau, M. A. (1985) *J. Bacteriol.* **163,** 806–811.
2. Bachmann, B. J., Low, K. B. & Taylor, A. L. (1976) *Bacteriol. Rev.* **40,** 116–167.
3. Bachmann, B. J. (1983) *Microbiol. Rev.* **47,** 180–230.
4. O'Farrell, P. H. (1975) *J. Biol. Chem.* **250,** 4007–4021.
5. Neidhardt, F. C., Vaughn, V., Phillips, T. A. & Block, P. L. (1983) *Microbiol. Rev.* **47,** 231–284.
6. Kiehn, E. D. & Holland, J. T. (1970) *Nature (London)* **226,** 544–545.
7. Sommer, S. S. & Cohen, J. E. (1980) *J. Mol. Evol.* **15,** 37–57.
8. Bravo, R. & Celis, J. (1984) in *Two-Dimensional Electrophoresis of Proteins*, eds. Celis, J. E. & Bravo, R. (Academic, New York), pp. 445–476.
9. Svedberg, T. & Petersen, K. O. (1940) *The Ultracentrifuge* (Clarendon, Oxford), pp. 406–407.
10. Wrinch, D. M. (1938) *Cold Spring Harbor Symp. Quant. Biol.* **6,** 122–134.
11. Bergmann, M. & Niemann, C. (1938) *Annu. Rev. Biochem.* **7,** 99–124.
12. Lloyd, D. J. & Shore, A. (1938) *Chemistry of the Proteins* (Blakiston, Philadelphia).
13. Schmidt, C. L. A. (1945) *The Chemistry of the Amino Acids and Proteins* (Thomas, Springfield, IL).
14. Cohn, E. J. & Edsall, J. T. (1943) *Proteins, Amino Acids and Peptides* (Reinhold, New York).
15. Goldberg, M. E. (1969) *J. Mol. Biol.* **46,** 441–446.
16. Wetlaufer, D. B. (1973) *Proc. Natl. Acad. Sci. USA* **70,** 697–701.
17. Wetlaufer, D. B. (1981) *Adv. Protein Chem.* **34,** 61–92.
18. Richardson, J. S. (1981) *Adv. Protein Chem.* **34,** 167–339.
19. Karplus, M. & Weaver, D. L. (1976) *Nature (London)* **260,** 404–406.
20. Go, M. (1981) *Nature (London)* **291,** 90–92.
21. Go, M. (1983) *Proc. Natl. Acad. Sci. USA* **80,** 1964–1968.
22. Hill, R. L., Delaney, R., Fellows, R. E. & Leboritz, H. E. (1966) *Proc. Natl. Acad. Sci. USA* **56,** 1762–1769.
23. Edelman, G. M., Cunningham, B. A., Gall, W. E., Gottlieb, P. D., Rutishauser, U. & Waxdal, M. J. (1969) *Proc. Natl. Acad. Sci. USA* **63,** 78–85.
24. Edelman, G. M. & Gall, W. E. (1969) *Annu. Rev. Biochem.* **38,** 415–466.
25. Sakano, H., Rogers, J. H., Hüppi, K., Brack, C., Traunecker, A., Maki, R., Wall, R. & Tonegawa, S. (1979) *Nature (London)* **277,** 627–633.
26. Honjo, T. (1983) *Annu. Rev. Immunol.* **1,** 499–528.
27. Parnes, J. R. & Seidman, J. G. (1982) *Cell* **29,** 661–669.
28. Williams, A. F. & Gagnon, J. (1982) *Science* **216,** 696–703.
29. Hood, L., Steinmetz, M. & Malissen, B. (1983) *Annu. Rev. Immunol.* **1,** 529–568.
30. Malissen, M., Minard, K., Mjolsness, S., Kronenberg, M., Goverman, J., Hunkapiller, T., Prystowsky, M. B., Yoshikai, Y., Fitch, F., Mak, T. W. & Hood, L. (1984) *Cell* **37,** 1101–1110.
31. Marchalonis, J. J., Vasta, G. R., Warr, G. W. & Barker, W. C. (1984) *Immunol. Today* **5,** 133–142.
32. Williams, A. F. (1984) *Immunol. Today* **5,** 219–221.
33. Marchalonis, J. J. & Barker, W. C. (1984) *Immunol. Today* **5,** 222–223.
34. Bränden, C.-I. (1980) *Q. Rev. Biophys.* **13,** 317–338.
35. Westhof, E., Altschuk, D., Moras, D., Bloomer, A. C., Mondragon, A., Klug, A. & Van Regenmortel, D. H. V. (1984) *Nature (London)* **311,** 123–126.
36. Tainer, J. A., Getzoff, E. D., Alexander, H., Houghten, R. A., Olson, A. J., Lerner, R. A. & Hendrickson, W. A. (1984) *Nature (London)* **312,** 127–134.
37. Gilbert, W. (1978) *Nature (London)* **271,** 501.
38. Doolittle, W. F. (1978) *Nature (London)* **272,** 581–582.
39. Blake, C. C. F. (1978) *Nature (London)* **273,** 267.
40. Blake, C. C. F. (1979) *Nature (London)* **277,** 598.
41. Blake, C. C. F. (1983) *Nature (London)* **306,** 535–537.
42. Südhof, T. C., Goldstein, J. L., Brown, M. S. & Russell, D. W. (1985) *Science* **228,** 815–822.
43. Gilbert, W. (1985) *Science* **228,** 823–824.
44. Burgoyne, L. A. & Skinner, J. D. (1982) *Nucleic Acids Res.* **10,** 665–673.
45. Williams, S. P., Athey, B. D., Muglia, L. J., Schappe, R. S., Gough, A. H. & Langmore, J. P. (1985) *Biophys. J.* **49,** 233–248.
46. Worcel, A., Strogatz, S. & Riley, D. (1981) *Proc. Natl. Acad. Sci. USA* **78,** 1461–1465.
47. Woodcock, C. L. F., Frado, L. L. Y. & Rattner, J. B. (1984) *J. Cell Biol.* **99,** 42–52.
48. Pettijohn, D. E. (1982) *Cell* **30,** 667–669.
49. Griffith, J. D. (1976) *Proc. Natl. Acad. Sci. USA* **73,** 563–567.