J. Mol. Evol. 15, 37–57 (1980)

Journal of
Molecular Evolution
© by Springer-Verlag · 1980

The Size Distributions of Proteins, mRNA, and Nuclear RNA

Steve S. Sommer and Joel E. Cohen

The Rockefeller University, 1230 York Avenue, New York, NY 10021, USA

Summary. The frequency distributions of size (molecular weight) and of numbers of subunits were determined from lists of over 500 mammalian and bacterial proteins. The size distribution of polypeptides is well fitted by a lognormal distribution with a median value of about 40,000 daltons and a deviation of 1.8. About 60% of all proteins exist in multimeric aggregates. Of the multimers 75% have either two or four subunits while less than 1% have an odd number of subunits that is greater than three. Over 90% of the time, a given multimer is composed of subunits of nearly equal size so that the size of a N-mer is lognormally distributed with a median value of N x 40,000 daltons and a deviation of 1.8. The distribution of polypeptide size and subunit number is similar for mammalian and bacterial proteins as well as for intracellular and extracellular proteins.

The sedimentation profiles of mRNA from HeLa and CHO cells indicate that the lengths of mammalian mRNA are lognormally distributed with a median value of 1.4 kb and a deviation of 2.0. This implies that, on the average, a mRNA species is only about 25% larger than the mature polypeptide it codes for. Therefore, at most a small fraction of mammalian mRNA could code for large precursor polypeptides which are then cleaved into a number of mature polypeptides (like polio mRNA), or for 3' coterminal mRNAs where the larger species contain the information for up to four proteins (like adenovirus mRNA).

The sedimentation profile of nascent nuclear RNA from HeLa suggests that the length distribution of transcription units has 2 components: An exponen-

This paper is dedicated to Harold Sommer

Offprint requests to: S.S. Sommer, Department of Pathology, Bldg. 10, Room 1AZ7, National Cancer Institute, Bethesda MD 20205

Abbreviations: hRNA – heterogeneous RNA
$L_{1/2}$ – in an exponential distribution, the increase in length required to reduce the frequency by a factor of 2
kb – kilobases
kd – kilodaltons
CHO cells – Chinese hamster ovary cells

0022–2844/80/0015/0037/$ 04.20

tial component that decays with a half-length of 10–15 kb, and a high frequency of very short molecules. However, other distributions (for example, the lognormal distribution) of transcription unit lengths could also be consistent with the data if one or more of the following occurred: Physiological cleavage of nascent chains, perturbation of non-rRNA transcription by actinomycin D, or degradation during isolation.

The length distribution of HeLa nuclear RNA labeled for 60 min is similar to that of nascent nuclear RNA, indicating that a completed hnRNA chain is quickly transported or degraded after being cleaved.

Key Words: Lognormal distribution – Subunit size – Mammalian protein – Bacterial protein – Sedimentation profile

Introduction

In this paper we describe the size distributions of proteins, mRNA, and nuclear RNA.

An important tool in our analysis is the lognormal distribution. The frequency curve of the lognormal distribution is positively skewed: It peaks early and then declines slowly. The lognormal distribution is so named because the logarithm of the variate plotted against the frequency has the shape of a Gaussian or normal curve. Just as a normal distribution can arise as the sum of many independent random quantities, the lognormal distribution can arise as the *product* of many independent random effects. A variate subject to a process of change will have a lognormal distribution if the change in the variate at each of the many steps in the process is a random proportion of the previous value of the variate (Aitchison and Brown, 1957). One might expect to find the lognormal distribution where exponential amplification exists; for example, where organisms divide or where wealth breeds more wealth.

The lognormal distribution appears often in biology and the social sciences. For example, the lognormal distribution is compatible with the frequency distribution of (1) the weight of human beings (Yuan, 1933), (2) the number of individuals in a species (Williams, 1937; Preston, 1948), (3) the number of viral lesions in plants infected with tobacco mosaic or bushy stunt viruses (Kleczkowski, 1949), (4) income in the U.S. (U.S. Dept. of Commerce, 1952), and (5) the number of inhabitants per town (Gibrat, 1931).

We find that the sizes of both proteins and mRNA are fit well by a lognormal distribution. In contrast, we tentatively find that both transcription units and partially processed hnRNA have exponential distributions of size.

We feel that the size distributions of proteins and RNA are of interest because they shed light on structure and on the overall organization of transcription and translation. They also eliminate any model of the evolution of proteins and RNA which does not predict the correct distributions.

Materials and Methods

(a) Cell Growth and Labeling Procedure. Suspension cultures of HeLa S3 cells (3–6 x 10^5 cells/ml) and Chinese hamster ovary cells were grown at 37°C in Eagle's medium

(Eagle, 1959) supplemented with 5% (v/v) fetal calf serum. The cells doubled every 24 h.

For determination of the size distribution of mRNA, cells were concentrated to 3 x 10^6 cells/ml and labeled with either carrier-free $H_3{}^{32}PO_4$ (200 μCl/ml) for 4 h in phosphate free medium, or 20 Ci/mmole (5- ^{3}H)-uridine (20 μCl/ml) for 2 h in regular medium.

For determination of the size distribution of nuclear RNA, HeLa cells were treated with 0.04 μg/ml actinomycin D for 25 min to suppress the transcription of rRNA (Perry, 1963). The cells were concentrated to 3 x 10^6 cells/ml and labeled with (5- ^{3}H)-uridine for either 30 s or 60 min. The cells were then poured over crushed frozen medium to stop incorporation rapidly (Derman and Darnell, 1974).

(b) Isolation of Nuclear RNA and mRNA. The cells were rinsed twice with isotonic buffer, swollen in hypotonic buffer, and then mechanically broken with a dounce homogenizer (Penman et al., 1963). The nuclei were pelleted, resuspended in hypotonic buffer, and vortexed 30 s with 0.1 volumes "magik" solution (1 volume 10% deoxycholate and 2 volumes 10% Tween 40). The nuclei were pelleted and the supernatant was pooled with the previous one to constitute the cytoplasmic fraction.

The nuclear fraction was extracted by a standard method which uses phenol at 65°C (Penman, 1966), and the cytoplasmic fraction was extracted with phenol at room temperature (Perry et al., 1972).

The poly A-containing cytoplasmic RNA was isolated by a modification of the method of Malloy et al. (1974). The RNA was layered onto poly U sepharose in 0.2 NETS (0.2 M NaCl, 10 mM EDTA, 10 mM Tris, 0.2% SDS), rinsed extensively with 20% formamide-ETS (10 mM EDTA, 10 mM Tris, and 0.2% SDS), and then the poly A-containing mRNA was eluted with 90% formamide-ETS.

(c) Sucrose Gradients. RNA samples were denatured by DMSO (dimethylsulfoxide) treatment. After ethanol precipitation samples were redissolved in 1 volume (usually 0.02 ml) of DMFO (dimethylformamide), 1 volume ETS, and 9 volumes DMSO. The solution was heated to 37°C, diluted 4 fold with ETS, and layered on a 15–30% sucrose gradient in 0.05 NETS (0.05 M NaCl,/10 mM EDTA,/10 mM Tris, 0.2% SDS) (Derman and Darnell, 1974).

An aqueous gradient was used rather than a DMSO gradient because a DMSO gradient has very poor resolving power in the bottom half of the gradient, and there was previous evidence that after denaturation with DMSO these aqueous gradients gave unaggregated profiles (Derman et al., 1976). Since it was important to verify that aggregation did not occur, aliquots of each nuclear RNA sample were sedimented in both aqueous and 99% DMSO gradients (Strauss et al., 1968). In each experiment the DMSO gradients verified that no aggregation occurred in the aqueous gradients. As an example, in one experiment the percentage of cpm that migrated faster than an internal 18S optical density marker was 82% for the aqueous gradient and 79% for the DMSO gradient. Likewise, the percentages of cpm migrating faster than 32S and 45S markers were 56% versus 55% and 33% versus 34% respectively. As a word of caution, nuclear RNA from Chinese hamster ovary cells does aggregate under the conditions of these aqueous gradients (M. Harpold, unpublished results).

(d) The Lognormal Distribution; Tests of Fit. A simple graphical test of the qualitative compatibility between an observed frequency distribution of molecular weight and the lognormal distribution is to plot the molecular weight on the ordinate and the percent of molecules (e.g. proteins) with a molecular weight less than or equal to the corresponding ordinate on the abscissa of logarithmic probability paper (Table 1). Logarithmic probability paper and its use are described by Aitchison and Brown (1957). The paper is designed so that if the observed frequency distribution is lognormal, the data will fall approximately along a straight line.

Let y [x] be the ordinate corresponding to the abscissa of x in logarithmic probability coordinates. If the data fall on a straight line, the median (m) is estimated as y(50%), and the deviation (d) is estimated as d = 1/2 [y(50%) / y(16%) + y(84%) / y(50%)] (Aitchison and Brown, 1957, p. 32). The logarithm of the deviation equals the standard deviation of the normal distribution produced by taking the logarithm of the variate. Therefore, 68% of molecules have molecular weights between m/d and m x d, and 95% have molecular weights between m/d^2 and m x d^2.

In all plots on logarithmic probability paper, we fitted a straight line through the data points by hand because the uncertainty in that procedure was small with these data. The deviations from the fitted lines were neither large nor systematic. Monte Carlo experiments showed that the statistical efficiency of estimating parameters using

Table 1. Example of how data are transformed for plotting on logarithmic probability paper[a]

Raw data			x[b]
Molecular weight of Subunit ($x10^{-3}$ daltons)	Number in each bin	No. in that bin and in smaller bins	(Percent of Proteins smaller than or equal to the largest member in the corresponding bin)
0– 9	2	2	1.0
10–19	12	14	6.8
20–29	21	35	17
30–39	47	82	40
40–49	41	123	60
50–59	34	157	76
60–69	21	178	86
70–79	8	186	90
80–89	8	194	94
90–99	5	199	96.1
109	1	200	96.6
130	1	201	97.1
140	2	203	98.1
142	1	204	98.6
155	1	205	99.0
160	1	206	99.5
165	1	207	100

[a] Actual data plotted in Fig. 1c

[b] x is plotted on the horizontal axis and the molecular weight of the largest member in the corresponding bin is plotted on the vertical axis of logarithmic probability paper

hand-fitted lines was not greatly inferior to more refined techniques when the distribution underlying the data was lognormal (Aitchison and Brown, 1957).

Where the molecular weights of individual molecules are known, it is possible to carry out numerical tests for departures from a normal distribution of the logarithms of the molecular weights. Such tests are equivalent to tests of departures from lognormality of the original molecular weights. Three such tests were applied here to mammalian and separately to bacterial intracellular protein subunit molecular weights: A test of skewness $b_1^{1/2}$, a test of kurtosis b_2, and the studentized range u = range/standard deviation (Pearson and Hartley, 1966, 1972). Each test was two-tailed, using the upper and lower 1% percentage points of the test statistics.

(e) The Exponential Distribution; A Graphical Test. A measured quantity (such as a molecular weight or length) has the exponential distribution if the probability that the quantity exceeds any value $x \geq 0$ is $e^{-\mu x}$. A graphical test of whether an observed frequency histogram (probability density function) conforms approximately to an exponential distribution is to plot the logarithm of frequency on the ordinate against the measured variate on the abscissa. The observed points approximate a straight line if they are exponentially distributed, and the slope of the line in such a plot estimates $-\mu$. In a plot where the abscissa is length, we define the "half length" $L_{1/2}$ to be the increase in length required for the frequency to fall by a factor of 2. $L_{1/2}$ equals the median value and $L_{1/2}/(\ln 2)$ equals the mean, where ln 2 is the natural logarithm of two.

(f) Conversion of Sedimentation Profile to Length Frequency Distribution. To convert the sedimentation profile of mRNA or nuclear RNA labeled for 60 min to a distribution of molar frequency versus length in nucleotides, the cpm in each fraction were divided by both the mid-fraction length and the range of lengths in that fraction. Division by length corrected for the fact that a mole of mRNA 2N nucleotides in length contributed twice as many cpm as a mole of mRNA N nucleotides in length. Division by the range of lengths corrected for the fact that a fraction near the bottom of the gradient had a much larger range of length than a fraction near the top of the gradient. For nuclear RNA labeled for 30 s it was unnecessary to divide by total length because all sizes of molecules were labeled at their 3' ends with the same average number of nucleotides.

The length of RNA was determined by $L = kf^{1.7}$, where L = length in nucleotides, k = a constant, and f = fraction number. This empirically derived formula gives a good estimate of length throughout the gradient (Derman et al., 1976). For mRNA, k was determined by using 4S, 18S, and 28S markers, and for nuclear RNA, k was determined by using 18S, 32S, and 45S markers.

For mRNA, the high molar frequencies found in the first 3 fractions (1-260 nucleotides) are discarded for three reasons. First, given a minimum length of 70 nucleotides for the poly(A), the maximum size of the protein the RNA could code is 6 kd. Second, tiny amounts of degradation of larger mRNA will greatly amplify the frequency in these fractions. Finally, an abundant species of RNA of length 150 nucleotides binds to poly U sepharose columns by virtue of having sequences complementary to mRNA (Jelinek and Leinwand, 1978).

The 30 s label of nuclear RNA labels nascent chains at their 3' ends. To estimate a frequency distribution of completed chains, we take minus the derivative of the nascent chain frequency distribution because: 1) completed chains of each length L produce a uniform distribution of nascent chains from N to L (N = length of RNA polymerized during the labeling period), and 2) completed chain transcripts of equal frequency but different lengths contribute to the nascent chain profile in proportion to their lengths. Consequently, the nascent distribution (see Fig. 5A) decreases monotonically for lengths greater than N. The decrease in an interval reflects the relative frequency of completed chains in that interval (for a more detailed discussion, see Derman et al., 1976). Since the differences between intervals in a nascent profile amplifies small errors, we take the derivative of the fitted curve as a more accurate way of estimating the completed chain distribution.

(g) Theory of Random Breakage of Molecules. Random breakage of molecules approximates the effect of thermal and endonucleolytic degradation of RNA. The following calculations were carried out to determine if an observed distribution of chain lengths which was approximately exponential could have arisen from random breakage of an initial distribution of molecular lengths which was radically different from exponential.

In the Appendix, it is shown that if molecules which are initially all exactly H units long are broken at random by a Poisson-distributed number of points, then the probability that a random fragment exceeds any length x, where x is strictly less than H, is $e^{-\lambda x} - (\lambda x e^{-\lambda x}) / (\lambda H + 1)$. In this formula λ is the average number of break points per unit of length. Except for a non-zero probability that a "fragment" will be of length H because no break points occurred on a molecule, this distribution of fragment lengths is very nearly exponential for the values of λ and H of interest here (see Results).

What effect would random breakage have on molecules whose initial sizes were lognormally distributed? We assumed that the average number of breaks would be proportional to the initial size of the molecule. We approximated a lognormal distribution of median 32 kb and deviation 4 by choosing 11 lengths equally spaced on a logarithmic scale: $H_k = 2^k$ kb, where k = 0, 1, ..., 10. To each value of H_k we assigned the probabilitiy density of such a length in a lognormal distribution (Table 2). (Technically speaking, we thus approximated the lognormal distribution as a discrete mixture of constant distributions.) We then divided the interval from 0 to 1024 kb into 512 equal

Table 2. Approximation to a lognormal distribution with median 32 kb and deviation 4

Length category k	Length (kb) H_k	Assigned probability P_k
0	1	.008
1	2	.028
2	4	.064
3	8	.122
4	16	.177
5	32	.201
6	64	.177
7	128	.122
8	256	.064
9	512	.028
10	1024	.008

intervals of length 2 kb and computed the probability contributed to each of these intervals by the fragments randomly broken from each initial chain of length H_k. For example, all fragments of the chains initially $H_0 = 1$ kb and $H_1 = 2$ kb long fell in the interval $0 < L \leqslant 2$, while fragments of chains initially $H_2 = 4$ kb long fell partly in the interval $0 < L \leqslant 2$ and partly in the interval $2 < L \leqslant 4$. The distribution among these intervals was determined using the formulas derived in the Appendix.

These calculations were carried out for 3 values of λ: $\lambda_8 = 1/8$, $\lambda_{32} = 1/32$ and $\lambda_{128} = 1/128$ (see Results).

Results

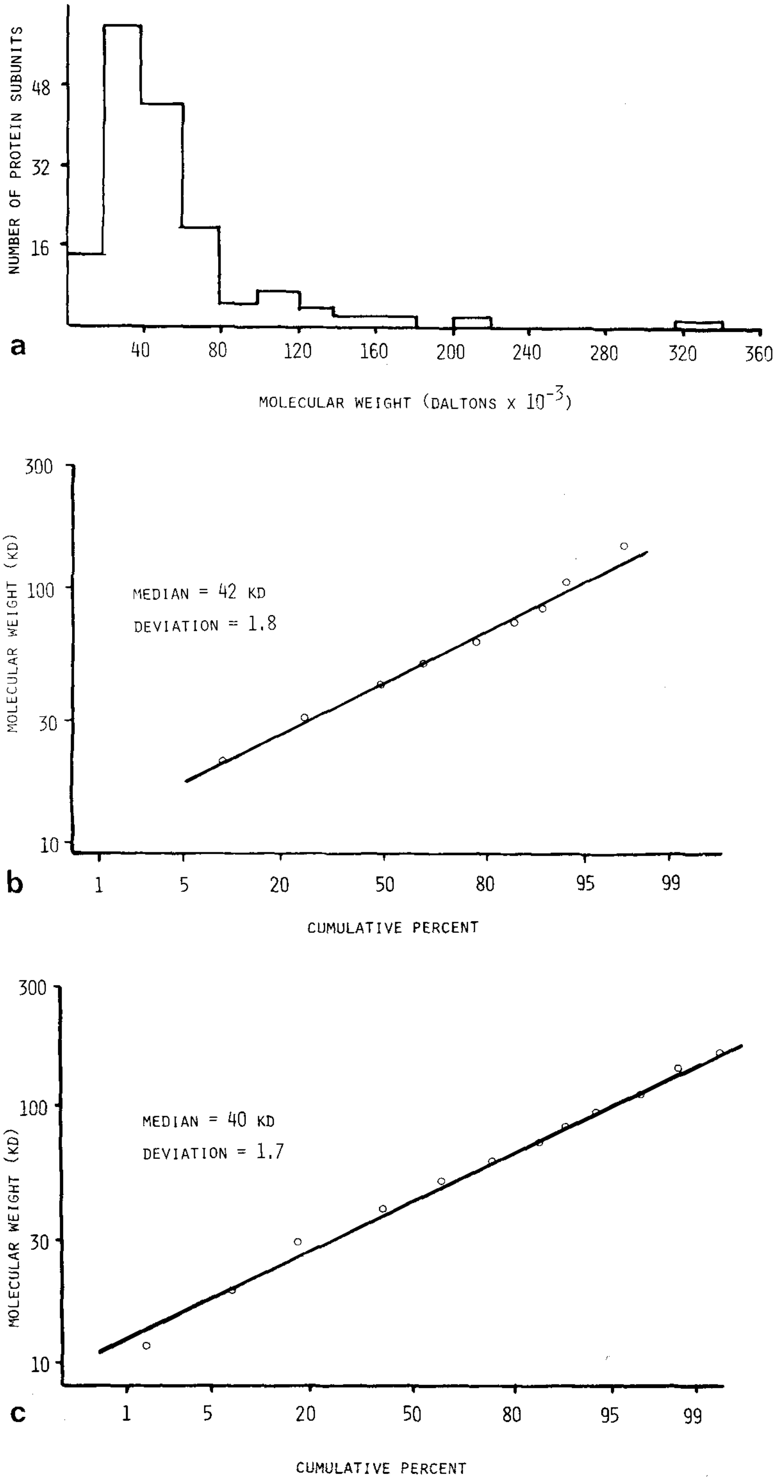
1) The Size Distribution of Proteins. Mammalian cells contain more than 10,000 different proteins (Bishop, 1974). The size distribution of these proteins cannot be accurately estimated by separating total cellular proteins according to molecular weight because, even in unspecialized cells like HeLa, the great abundance of a few proteins obscures the general pattern (Peterson and McConkey, 1976). An alternative approach is to plot the molecular weights of the hundreds of proteins that have been purified. Using a comprehensive list of the molecular weights of over 500 multimeric proteins and their subunits (Darnall and Klotz, 1976), the size distribution for intracellular protein subunits from mammalian cells was determined (Fig. 1a). The distribution peaks early and then declines slowly. Graphing the data on logarithmic probability paper indicates that a lognormal distribution fits the data well (Fig. 1b). In addition, the three numerical tests fail to detect a significant departure from lognormality at the 2% level. We estimate a median of m = 42,000 daltons and a deviation of d = 1.8.

To test whether these same data are compatible with other common distributions, we matched the median and variance estimated from the lognormal distribution to normal, uniform and exponential distributions. In each of these three distributions, a substantial fraction of the molecules would have to have negative molecular weight in order for the distribution to have the median and variance estimated from the data. This biologically nonsensical result argues against these alternative distributions.

Moreover, the points in Fig. 1b are nearly linear over a 10-fold range of molecular weight. The normal, uniform, and exponential distributions all show a marked curvature on logarithmic probability paper over a 10-fold range of molecular weight (Fig. 2).

When the subunit molecular weights of intracellular bacterial proteins are graphed on logarithmic probability coordinates (Fig. 1c), the fitted lognormal distribution has approximately (within 5%) the same median and deviation found with mammalian proteins. The deviations from linearity are minor and not significant according to two of the three numerical tests, b_1 and u. However, b_2 is significantly high. This apparent departure from lognormality could arise if a few of the published subunit molecular weights were not the weights of the true minimal subunits. On the basis of the close similarity of the size distributions of the mammalian and bacterial intracellular protein subunits, we accept the lognormal distribution as a good description of the size of bacterial subunits as well.

The distributions of molecular weights of dimers and tetramers are fitted graphically by lognormal distributions of median value 88,000 daltons, deviation 1.8 and median value 160,000 daltons, deviation 1.7, respectively (Fig. 1d). The increasing median and


a
NUMBER OF PROTEIN SUBUNITS
48
32
16
40
80
120
160
200
240
280
320
360
MOLECULAR WEIGHT (DALTONS x 10^{-3})
b
300
100
30
10
MOLECULAR WEIGHT (KD)
MEDIAN = 42 KD
DEVIATION = 1.8
1
5
20
50
80
95
99
CUMULATIVE PERCENT
c
300
100
30
10
MOLECULAR WEIGHT (KD)
MEDIAN = 40 KD
DEVIATION = 1.7
1
5
20
50
80
95
99
CUMULATIVE PERCENT

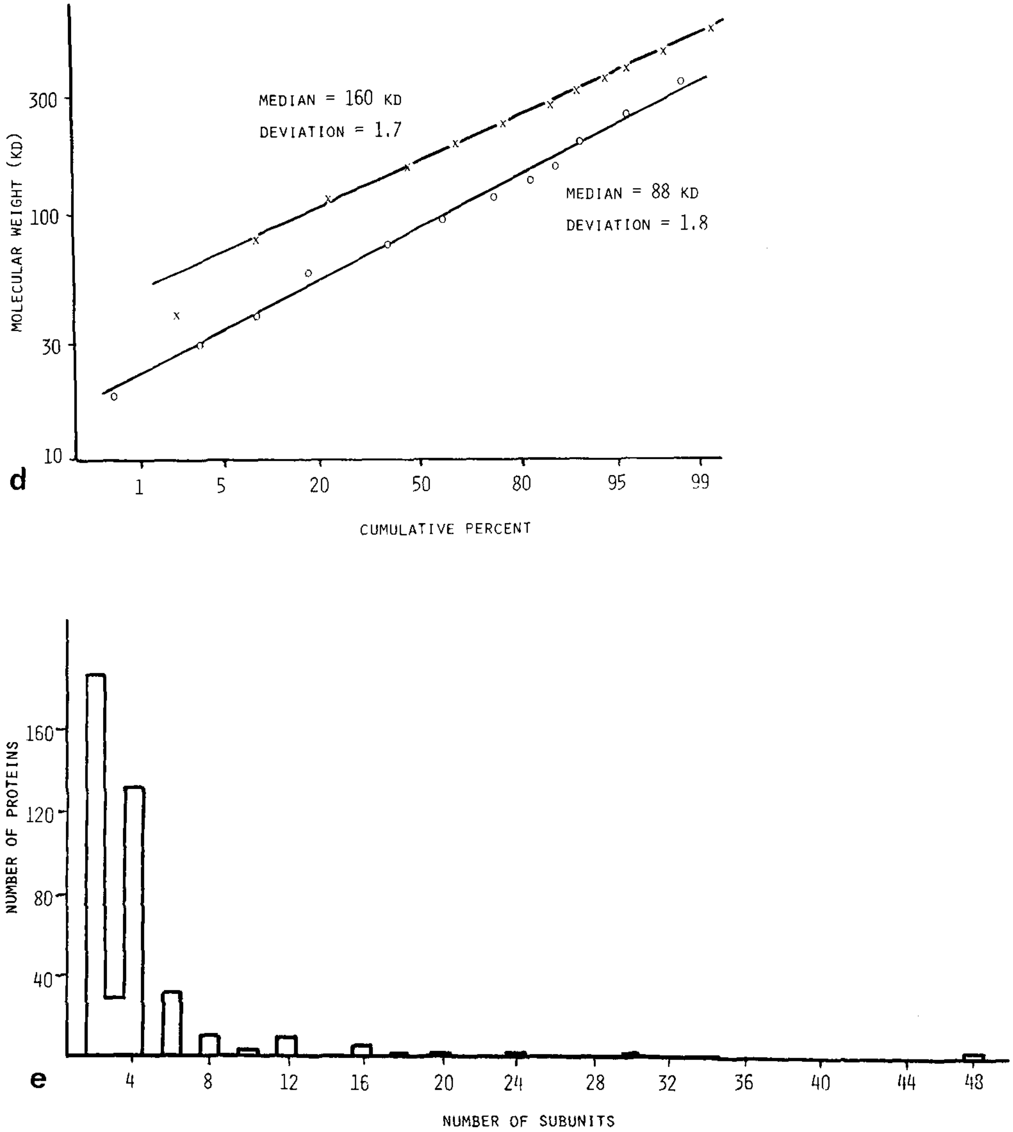


Fig. 1 a-e. The size distribution of proteins. The source of the data is Darnall and Klotz (1976). In b-d, the y-axis is molecular weight and the x-axis is the percent of proteins with less than or equal that molecular weight. When a given protein had more than one molecular weight due to divergent estimates, or isolates from different organisms, one estimate was chosen using a random number table. All lines were fitted by hand.

a. Size distribution of 153 intracellular mammalian protein subunits.
b. The data in a plotted on logarithmic probability paper.
c. The size distribution of 207 subunits of intracellular bacterial proteins.
d. Size distributions of 180 dimeric (-0-) and 140 tetrameric (-X-) proteins.
e. Distribution of subunit number from 439 proteins.

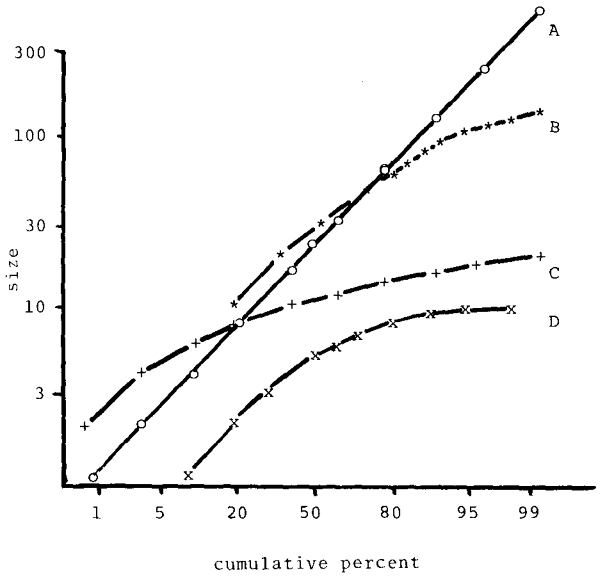


Fig. 2. The appearance of a (A) lognormal, (B) exponential, (C) normal, and (D) uniform distribution on logarithmic probability coordinates. For a very wide range of values of the mean or variance, all of these distributions but the lognormal show a marked curvature over a one order of magnitude range of sizes.

the nearly constant deviation reflect the fact that over 90% of the time, the molecular weights of the subunits in a given multimer are within 20% of one another.

Figure 1e shows the distribution of subunit number for all multimeric proteins. Three-fourths of the multimers are either dimers or tetramers. The number of subunits varies by more than a factor of 20. Over 90% of the multimers have an even number of subunits. Over 90% of the multimers with an odd number of subunits are trimers.

The distributions of subunit number for both mammalian intracellular and bacterial intracellular multimers are very similar to the distribution for all multimers (data not shown).

How does the frequency of monomeric proteins compare to the frequency of multimers? Unfortunately, no comprehensive list of molecular weights of monomers exists to complement the list by Darnall and Klotz (1976). However, the percentage of monomeric proteins can be estimated if it is assumed that the mean molecular weight of monomeric proteins is close to that of subunits of multimeric proteins. The method of estimation we use depends only on the mean molecular weights and not on the distribution around the means. If M is the fraction of all proteins that are monomers, mw_{mono} is the mean molecular weight of monomeric proteins, mw_{multi} is the mean molecular weight of multimeric proteins, and mw_{total} is the mean molecular weight of all proteins, then these quantities satisfy the equation

$$M(mw_{mono}) + (1-M)mw_{multi} = mw_{total},$$

which we now solve for an estimate of M.

For mammals, the mean size of 153 protein *subunits* is 51,500 daltons. For bacteria, the mean size of 207 protein subunits is 49,600 daltons. The difference between these means is not statistically significant. The mean weight of 50,400 daltons of the combined 360 protein subunits estimates the mean molecular weight of monomeric proteins.

For mammals, the mean size of 133 multimeric proteins is 215,900 daltons. For bacteria, the mean size of 175 multimeric proteins is 240,400 daltons. Again, the difference between these means is not statistically significant. The mean weight of the combined 308 multimeric proteins is 229,800 daltons.

Finally, the mean size of 77 human plasma proteins (extracellular proteins from a list by Masson, 1976) is 158,700 daltons and of 95 metalloenzymes (a mixture of bacterial and mammalian intracellular proteins from a list by Vallee and Wacker, 1976) is 158,100. The difference between these means is again not statistically significant. The mean weight of the combined 172 proteins is 158,400 daltons.

Using the above equation, the fraction M of proteins that are monomers may be estimated as:

$$M(50{,}400) + (1-M)(229{,}800) = 158{,}400, \text{ hence } M = 0.40$$

2) Size Distribution of mRNA. How does the size of mRNA compare to the size it must have to code for protein subunits? Unfortunately, the size distribution cannot be accurately determined from specific cellular mRNAs because too few have been isolated. However, unlike the proteins the majority of the mass of mRNA in a cell like HeLa is divided among hundreds of different species (Bishop et al., 1974). Therefore the sedimentation profile of cellular mRNA can be used to estimate its size.

The transformation of the sedimentation profile to the molar frequency size distribution is described in Materials and Methods. The size distribution is well described by a lognormal distribution with a median length of approximately 1.4 kb and a deviation of 2.0 (Figs. 3a and 4a). Likewise, the size of mRNA isolated from Chinese hamster ovary cells is fitted by a lognormal distribution with approximately (within 5%) the same median and deviation (Figs. 3b and 4b). Although numerical tests are not possible here, the graphical test indicates that the description of the data by a lognormal distribution is, if anything, even better than for the mammalian protein subunits (Fig. 1b), where numerical tests detected no departures from lognormality.

To determine the maximal coding capacity of mRNA, we estimate the average molecular weight of amino acids in proteins at 110 daltons. This figure was obtained in each of 3 different ways: 1) from the amino acid composition of total KB cell (human) protein, 2) from the amino acid compositon of total *E. coli* protein, and 3) from the average composition of 30 proteins selected randomly from a compilation of amino acid compositions of purified proteins (Polasa and Green, 1967; Sueoka, 1961; Reeck, 1976). We shall assume that the efficiency of translation, as measured by the number of initiations per mRNA molecule per unit time, is independent of the size of mRNA. Taking 1.4 kb as the median length of a lognormal distribution of mRNA, 2.0 as the deviation, and 100 nucleotides as the average length of poly A (Puckett and Darnell, 1976; Sawicki et al., 1977), mRNA could, at most, produce a distribution of polypeptide chains with a median molecular weight of 51,000 daltons and a deviation of 2.0.

Since the observed distributions of polypeptide size have a median on the order of 40,000 daltons and a deviation of 1.8, 3/4 of the length of a typical mRNA molecule

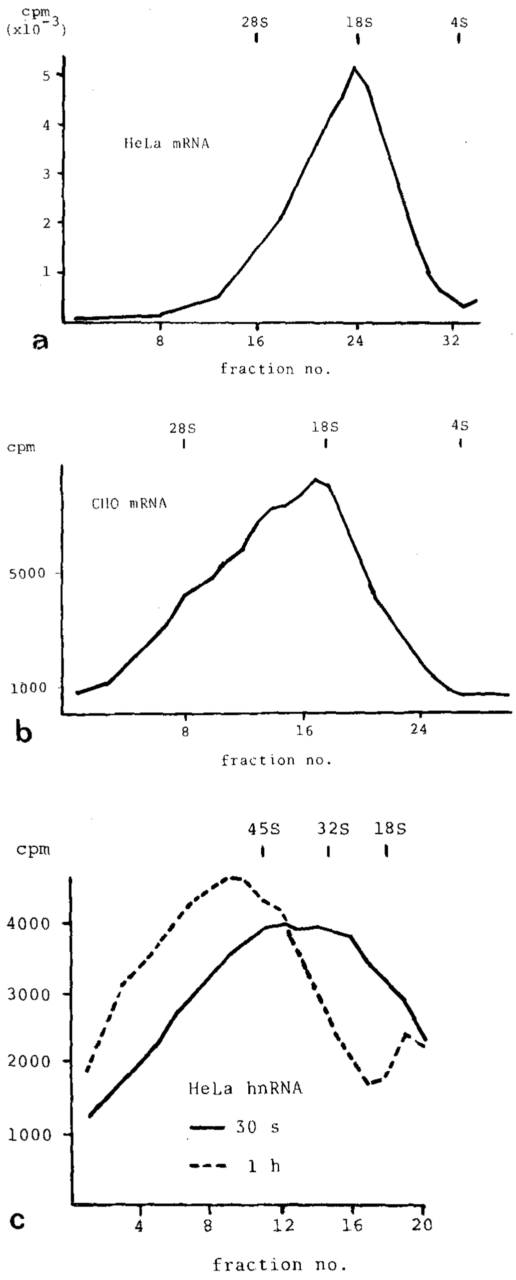


Fig. 3a-c. Sucrose gradient profiles of mRNA and nuclear RNA.

a. 15–30% sucrose gradient of mRNA from HeLa cells labeled with $H_3{}^{32}PO_4$ for 6 h. The gradient contains poly A$^+$ mRNA but the results also apply to poly A$^-$ because it has a similar if not identical sedimentation profile (Milcarek et al., 1974).

b. 5–20% sucrose gradient of mRNA from CHO cells labeled 2 h with ^{3}H-uridine.

c. 15–30% sucrose gradient of HeLa cell nuclear RNA labeled with ^{3}H-uridine for 30 s (–) and 1 h (- - -), respectively

is translated. Since the smaller mRNA species are too small to code for the larger polypeptides, they must code for the smaller polypeptides. Therefore to produce a lognormal distribution of protein size, the larger mRNA molecules must, in general, code for the larger polypeptides.

We emphasize that these conclusions only apply to the majority of mRNA species. If a minority behaved aberrantly, they would not be detected by these methods.

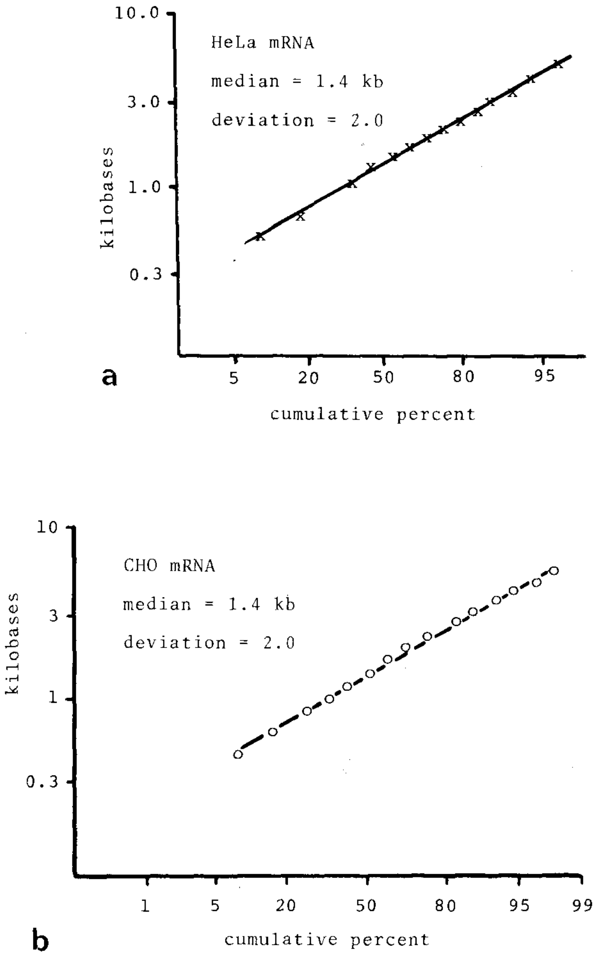


Fig. 4. The size distribution of mRNA from (a) HeLa and (b) CHO cells plotted on logarithmic probability coordinates

3) Size Distribution of Transcription Units in HeLa Cells. To compare the size distribution of transcription units to that of mRNA, HeLa cells were labeled with ^{3}H-uridine for 30 s. The nuclear RNA was extracted and then sedimented through a sucrose gradient under conditions that prevent aggregation (see Materials and Methods). The sedimentation profiles (Fig. 3c) were converted to graphs of frequency versus length in kb. The nascent chain size distribution (Fig. 5a) has 2 components: (1) an exponential component with a half-length in the range of 10–15 kb (based on 3 different experimental determinations), and (2) a high frequency of very short molecules.

The frequency of very short molecules is probably a conservative estimate. Based on the maximal elongation rates of HeLa 45S pre-rRNA, polio, *Chironomus tetans*, and *E. coli* RNA polymerases, the elongation length E during the 30 s label is likely to be greater than 0.3 kb and possibly as much as 3 kb (Greenberg and Penman, 1966; Darnell et al., 1967; Egyhasi, 1975; Bremer and Yuan, 1968). Therefore, the true frequency of a molecular species of length L which is shorter than E is E/L because, in 30 s, E/L completed copies of the molecule are transcribed.

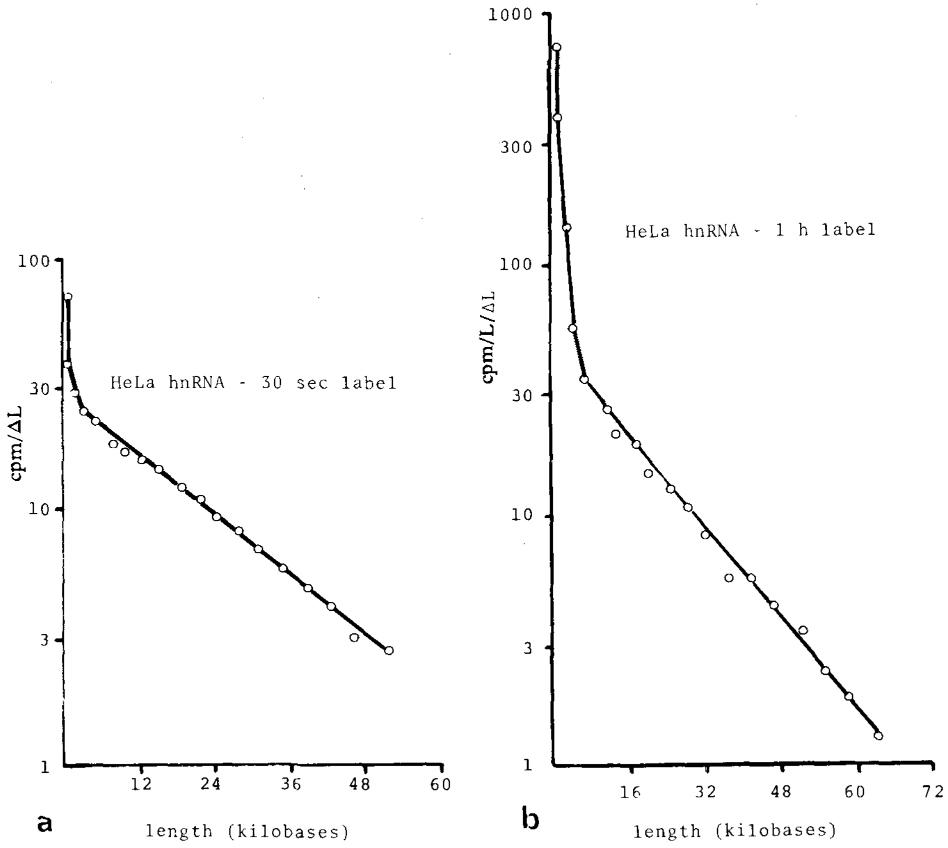


Fig. 5. The size distribution of HeLa hnRNA labeled for (a) 30 s and (b) 60 min with ^{3}H-uridine plotted on semi-logarithmic coordinates

The observed distribution reflects the actual size of transcription units (distance from the initiation to the termination of RNA polymerase) if there is not physiological cleavage of nascent chains, no perturbation introduced by shutting off rRNA synthesis with actinomycin D, and no degradation during the isolation of nuclear RNA. However, there may well be such cleavage, perturbation, and degradation. Nascent chains are cleaved during the transcription of *E. coli* rRNA and adenovirus late nuclear RNA (Nikolaev et al., 1973; Nevins and Darnell, 1978). Actinomycin D, at 100 times the concentration, perturbs hnRNA processing in human and *Drosophila* cells (Herman and Penman, 1977; Levis and Penman, 1975). Finally, degradation during isolation is always a potential problem because of the high concentrations of RNAses. However, the likelihood of degradation is difficult to evaluate because it is not known whether any isolation procedure for mammalian nuclear RNA quantitatively preserves molecules in the range of 15–100 kb.

The precise effects of processing and actinomycin D depend on their mechanisms of action. However, thermal and nucleolytic degradation, which cut randomly in

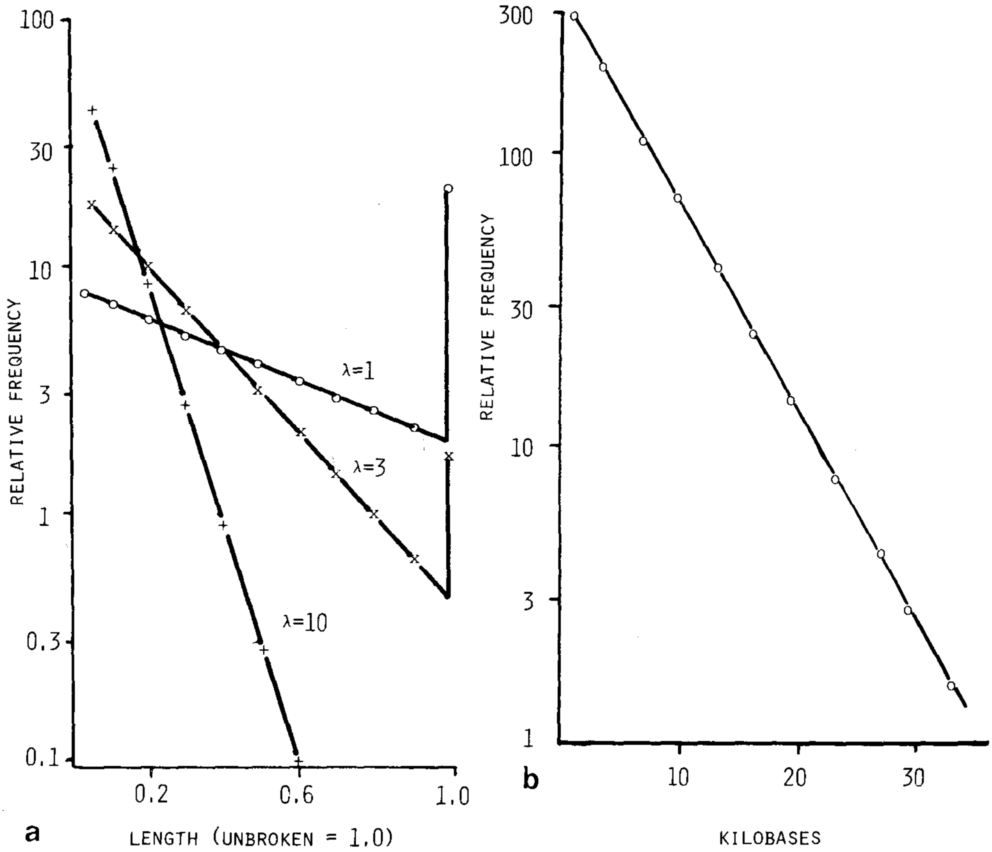


Fig. 6a, b. Frequency distribution of fragment length of chains randomly broken at a random number of points. With the scaling of the coordinates used here, a straight line indicates that the frequency distribution resembles an exponential distribution.

a. Initially all chains are of equal length. When an average of λ = 1 break point occurs per initial chain, the distribution of lengths is exponential with a marked discontinuity at 1 unit. As λ progresses through 3 and 10 break points per initial chain, the discontinuity decreases until it is negligible.

b. A nearly exponential distribution of chain length resulting from a lognormally distributed population subjected to random breaks. Initially, chains are approximately lognormally distributed with median 32 kb and deviation of 4. When a break point occurs, on average only once every 128 kb, breakage is so rare that the resulting distribution of fragments resembles the initial lognormal distribution. When a break point occurs on average every 32 kb, the frequency distribution is intermediate between lognormal and exponential. When a break point occurs on average every 8 kb (as shown in figure), hardly any trace of the initial lognormal distribution survives and the distribution of fragment length is very close to exponential

proportion to size, could convert unique and lognormal distributions into approximately exponential distributions (Fig. 6). Therefore, if degradation occurred, a variety of distributions of transcription unit size would appear to be exponential.

4) Size Distribution of Accumulated Nuclear RNA. The profile of nuclear RNA labeled for 60 min is similar to the completed chain profile. The half-length of the exponential component is 13 kb, indicating that a completed hnRNA chain is quickly transported or degraded after being cleaved. This agrees with previous conclusions from UV inactivation experiments in human and mouse cells, and pulse chase experiments in *Drosophila* cells (Goldberg et al., 1977; Giorno and Sauerbier, 1976; Levis and Penman, 1977).

The observed distribution only reflects the real distribution if actinomycin D does not appreciably perturb transcription and processing, and if the RNA is not degraded during isolation.

Discussion

a) Proteins. Our analysis of protein sizes and subunit numbers indicates five conclusions. The size of polypeptides is fit well by a lognormal distribution with a median value of 40,000 daltons and a deviation of 1.8. About 60% of all proteins exist in multimeric aggregates. 75% of the multimers have either 2 or 4 subunits while less than 1% have an odd number of subunits that is greater than 3. Over 90% of the time, a given multimer is composed of subunits of nearly equal size so that the size of an N-mer is lognormally distributed with a median value of N x 40,000 daltons and a deviation of 1.8. The distribution of polypeptide size and subunit number is similar for mammalian and bacterial proteins as well as for intracellular and extracellular proteins.

These conclusions require caution because the proteins that have been isolated are, in general, the more abundant proteins. If the size of proteins depends on their abundance, the sample available will not be representative of protein sizes as a whole. Since the mRNA and nuclear RNA sampled are predominantly the abundant molecules, the same caveat applies to generalizations about their sizes.

Our results may be compared with those of previous studies on the sizes of proteins.

The 360 intracellular mammalian and bacterial protein subunits included in Figs. 1a and 1c have a mean molecular weight of 50,419 daltons and a standard deviation of 32,793 daltons. Hopkinson et al. (1976) find the molecular weights of 99 enzyme subunits studied electrophoretically in man to have a mean and standard deviation of 45,798 ± 20,699 daltons. Edwards et al. (1977) find the molecular weights of soluble "non-enzyme" protein subunits from human autopsy tissues to have a mean and standard deviation of 54,600 ± 42,900 daltons. They also find 134 vertebrate enzymes from an earlier list by Darnall and Klotz to have a mean subunit molecular weight of 53,400 daltons, and 35 vertebrate "non-enzymes" to have a mean molecular weight of 51,700 daltons. Nei et al. (1976) find the molecular weights of 119 mammalian protein subunits taken from an earlier list by Darnall and Klotz to have a mean and standard deviation of 45,102 ± 24,531 daltons. Thus all these studies indicate a mean protein subunit molecular weight near 50,000 daltons and a standard deviation near 30,000 daltons.

We find that the frequency distribution of subunit molecular weight is well described by a lognormal distribution. Hopkinson et al. (1976) and Edwards et al. (1977) give frequency histograms of subunit molecular weight but do not fit any frequency law to these histograms. Qualitatively their histograms have the skewness characterisitic of the lognormal distribution. Nei et al. (1976) successfully fit a gamma distribution to their frequency histogram of 119 mammalian protein subunit sizes but offer no rationale for the choice of the gamma distribution. Although we have shown that several distributions other than the gamma cannot describe our data on subunit sizes, neither we nor Nei et al. (1976) have performed a comparative test to see whether the lognormal distribution or the gamma distribution describes the size data better. This question remains open.

We estimate that 40% of all proteins are monomers. Hopkinson et al. (1976) find 28 monomers in their sample of 100 enzymes. Among proteins with more than one subunit, we find that approximately 40% are dimers and approximately 68% have two or four subunits. The corresponding figures derived from the smaller sample of Hopkinson et al. (1976) for multimeric human enzymes are 60% and 93%.

We find that the mean protein subunit size is very similar in eukaryotes and prokaryotes, in intracellular and in extracellular proteins. Hopkinson et al. (1976) and Edwards et al. (1977) find a mean protein subunit size in humans comparable to that in vertebrates. Hence it seems likely that the overall means for mammalian and for bacterial protein subunits do not disguise a large variation from species to species, but that the mean molecular weight of protein subunits in individual species varies only slightly from the overall mean. This likelihood requires confirmation by the study of the protein subunit size distribution in other individual species.

One should gain insight into the forces governing protein size by finding systems where protein size differs. If temperature affects size then thermophilic organisms should have a different distribution. If protein size can be reduced, the size of proteins from small virions might be smaller because these organisms are under strong selective pressure to minimize the size of their genetic information, as judged by the out of phase genes in ϕX174 and SV40 (Sanger et al., 1976; Reddy et al., 1978).

The implications of the variation in protein subunit molecular weight for heterozygosity in natural populations of organisms are explored by Nei et al. (1976, 1978) and Koehn and Eanes (1978).

b) mRNA. The sedimentation profiles of mRNA from HeLa and CHO cells indicate that the lengths of mammalian mRNA are fit well by a lognormal distribution with a median value of 1.4 kb and a deviation of 2.0. This implies that the majority of mRNA is monocistronic because, on the average, a mRNA species in only about 25% larger than the mature polypeptide it codes for. Therefore, at most a minority of mammalian mRNA could code for large precursor polypeptides which are then cleaved into a number of mature polypeptides (like polio mRNA; Villa-Komaroff et al., 1975) or for 3' coterminal mRNAs where the larger species contain the information for up to 4 proteins (like adenovirus mRNA; Nevins and Darnell, 1978).

The deviation of mRNA size is somewhat greater than the deviation of polypeptide size (2.0 versus 1.8). This suggests that the length of the noncoding region is not closely correlated with the length of the coding region. In other words, while the average noncoding region is roughly 25% of the coding region, some noncoding regions are smaller and some are larger.

The size of mRNA from a wide variety of eukaryotic cells is quite similar. For example, sedimentation of ^{3}H-labeled mosquito mRNA with co-extracted ^{14}C-labeled human mRNA suggests that the mRNA size distribution in both species is virtually identical (Spradling et al., 1974). Although other published sedimentation profiles do not have internal controls and vary in their accuracy, it is safe to say that mRNA from paramecium, sea urchin, and fruitfly are close (at most within a factor of 2) in size to mammalian mRNA (Hruby et al., 1977; Nemer et al., 1975; Levis and Penman, 1977).

The data from the three cellular mRNA molecules that have been sequenced are compatible with the conclusions drawn by analyzing sucrose gradients. In rabbit α and β globin and chicken ovalbumin mRNA, the percentages of sequences in the nontranslated region are 20%, 22%, and 39%, respectively, with an average value of 27% (Proudfoot et al., 1977; Proudfoot, 1977; Baralle, 1977; Efstratiadis et al., 1977; MacReynolds et al., 1978).

It will be of interest to determine if the size distribution of mRNA differs in some organisms. Since bacteria contain polycistronic mRNAs derived from operons, the size distribution of their mRNA may be different. Since *Acetabularium* can live for months and even regenerate its cap without a nucleus (Brachet, 1967), its preponderance of post-transcriptional controls may be reflected in mRNA with a different size distribution.

c) Nuclear RNA. The sedimentation profile of nuclear RNA labeled for 30 s and 60 min suggests that the length distribution of both transcription units and accumulated nuclear RNA has an exponential component with a half-length of about 10–15 kb, and a high frequency of very short molecules.

Derman et al. (1976) reported that, neglecting the very small nuclear RNA, half the nuclear RNA in HeLa cells is synthesized from transcription units of less than 5 kb. When their data are plotted on semi-logarithmic paper, the curve is similar to the curve in Fig. 5a, but the half-length is only 6 kb. It is very unlikely that the discrepancy is due to aggregation of the RNA in our experiments because the RNA did not decrease in size when sedimented on a parallel 99% DMSO denaturing gradient. Since speed is crucial for the isolation of the larger molecules, perhaps Derman et al. got smaller RNA because they extracted 4 samples concurrently, while we extracted only 2 samples concurrently. In one experiment where 4 samples were extracted together we also observed smaller RNA.

If the distribution of hnRNA transcription unit length is exponential, then processing converts an exponential distribution of primary transcripts to a lognormal distribution of mRNA. However, since more than 90% of the primary hnRNA transcripts are larger than the median size of mRNA, it is still possible that a segment of each primary transcript gives rise to a single mRNA molecule.

So far as we know, detailed models have not yet been proposed which predict quantitatively the size distributions of hnRNA, mRNA, and proteins established in this paper. These facts constrain future models for the evolution of these classes of molecules and serve as a challenge to construct testable models which will explain them.

Appendix

Random Breakage at a Random Number of Points

Consider a population of molecules which are initially all exactly H units long. Suppose each molecule is broken at random at N ponits, where the number N of break points varies randomly according to a Poisson distribution with mean λH. We write the mean number of break points as λH so that λ describes the average number of breaks per unit of length. Thus if the initial population of molecules were twice as long as the population we are considering, we would expect on average twice as many breaks per molecule in this model. Let L be the length of a random fragment of a molecule produced by this process. This Appendix calculates the cumulative distribution function (cdf) of L. The cdf of L is denoted $F(x) = P(L \leqslant x)$ and is defined as the probability that the length of a random fragment is less than or equal to x.

We use notation that is standard in the theory of stochastic processes (Karlin and Taylor, 1975).

By assumption, the probability that there are $n = 0, 1, 2, \ldots$ break points on an initial molecule is $e^{-\lambda H} (\lambda H)^n/n!$. If $N = n$, there are $n + 1$ fragments with probability 1. So the expected number of fragments if $\Sigma_{n=0}^{\infty} (n + 1)e^{-\lambda H} (\lambda H)^n/n! = \lambda H + 1$, and the probability that a random fragment arises from a molecule broken at $ne^{-\lambda H}$ points is $P(N = n) = (n + 1)\, e^{-\lambda H}(\lambda H)^n/[n!\, (\lambda H + 1)]$. This differs from a Poisson distribution because the more points at which an initial molecule is broken, the more random fragments it creates.

If $N = 0$, then $L = H$; that is, by chance, the original molecule remains unbroken. We define $C(s) = 1$ if $s < 1$, $C(s) = 0$ if $s \geqslant 1$. Then the probability that L exceeds x, given $N = 0$ is $P(L > x \mid N = 0) = C(x/H) = 1$ if $x < H$, $= 0$ if $x \geqslant H$.

If $N = n > 0$, then for any x such that $0 \leqslant x \leqslant H$, $P(L > x \mid N = n) = (1-x/H)^n$, (Feller, 1966).

The cdf is now obtained by combining these results: $1 - F(x) = P(L > x) = \Sigma_{n=0}^{\infty} P(L > x \mid N = n)P(N = n)$. Thus $P(L > x)(\lambda H + 1) = C(x/H)e^{-\lambda H} + e^{-\lambda H} \Sigma_{n=1}^{\infty} (n + 1)[\lambda H(1-x/H)]^n/n! = e^{-\lambda x}[\lambda H(1-x/H) + 1 + e^{-\lambda H(1-x/H)}(C(x/H)-1)]$. It is readily checked that $P(L > 0) = 1$ and $P(L > H) = 0$ as desired so that the probability mass is concentrated on $(0, H)$. For $x < H$, the formula simplifies: $P(L > x \mid x < H) = e^{-\lambda x} - (\lambda x e^{-\lambda x})/(\lambda H + 1)$. For any a, b such that $0 \leqslant a < b < H$, we use this formula to find $P(a < L \leqslant b) = P(L > a \mid a < H) - P(L > b \mid b < H)$, while if $b = H$, $P(a < L \leqslant b = H) = (P(L > a \mid a < H) - 0$.

This procedure was used to find the probability mass of L in the 20 intervals $(0.05(j-1), 0.05j]$, $j = 1, \ldots, 20$, for $H = 1$, $\lambda = 1, 3$ and 10 (Fig. 6a).

Acknowledgments. We gratefully acknowledge the information supplied by Marianne Salditt-Georgieff. We thank J.E. Darnell and Igor Tamm for helpful discussions and Walter Fitch and a referee for useful references.

This work was partially supported by National Science Foundation grant DEB74-13276.

References

Aitchison, J, Brown, J.A.C. (1957): The lognormal distribution, p. 102. Cambridge: Cambridge University Press

Baralle, F.E. (1977). Cell **10**, 549–558

Bishop, J.O. (1974). Cell 2, 81–86

Bishop, J.O., Morton, J.G., Rosebach, M., Richardson, R.M. (1974). Nature **250**, 199–204

Brachet, J. (1967). Nature **213**, 650–655

Bremer, H., Yuan, D. (1968). J. Mol. Biol. **38**, 163–180

Darnall, D.W., Klotz, I.M. (1976). In: CRC Handbook of biochemical and molecular biology: Proteins. Fasman, G.D., ed., Vol. **2**, pp. 325–371, Cleveland: CRC Press

Darnell, J.E., Girard, M., Baltimore, D., Summers, D.F., Maizel, J. (1967). In: Molecular biology of viruses. Cotter, J., ed., New York: Academic

Derman, E., Darnell, J.E. (1974). Cell **3**, 255–264

Derman, E., Goldberg, S., Darnell, J.E. (1976). Cell **9**, 465–472

Eagle, H. (1959). Science **130**, 432–437

Edwards, Y.H., Hopkinson, D.A., Harris, H. (1977). Ann. Hum. Genet. **40**, 267–277

Efstratiadis, A., Kafatos, F.C., Maniatis, T. (1977). Cell **10**, 571–586

Egyhazi, E. (1975). Proc. Nat. Acad. Sci. **72**, 947–950

Feller, W. (1966). An introduction to probability theory and its applications., Vol. **2**, New York: Wiley

Gibrat, R. (1931). Les Inégalitiés Economique, Paris: Libraire de Recueil, Sirey

Giorno, R., Sauerbier, W. (1976). Cell **9**, 775–786

Goldberg, S., Schwartz, H., Darnell, J.E. (1977). Proc. Nat. Acad. Sci. **74**, 4520–4523

Greenberg, H., Penman, S. (1966). J. Mol. Biol. **21**, 527–535

Herman, R.C., Penman, S. (1977). Biochemistry **16**, 3460–3465

Hopkinson, D.A., Edwards, Y.H., Harris, H. (1976). Ann. Hum. Genet. **39**, 383–411

Hruby, P.E., Maki, R.A., Cummings, D.J. (1977). Biochim. Biophys. Acta, **47**, 89–96

Jelinek, W., Leinwand, L. (1978). Cell **15**, 205–214

Karlin, S., Taylor, H.M. (1975). A first course in stochastic processes. New York: Academic

Kleczkowski, A. (1949). Ann. Appl. Biol. **36**, 139–152

Koehn, R.K., Eanes, W.F. (1978). Evolutionary Biol. **11**, 39–100

Levis, R., Penman, S. (1977). Cell **11**, 105–113

MacReynolds, L.A., O'Malley, B.W., Nesbet, A.D., Fothergill, J.E., Givol, D., Fields, S., Robertson, M., Brownlee, G.G. (1978). Nature **273**, 723–728

Malloy, G.R., Jelinek, W., Salditt, M., Darnell, J.R. (1974). Cell **1**, 43–53

Masson, P.L. (1976). In: CRC Handbook of biochemical and molecular biology: Proteins. Fasman, G.D., ed., Vol. **2**, pp. 242–253, Cleveland: CRC Press

Milcarek, C., Price, R., Penman, S. (1974). Cell **3**, 1–10

Nei, M., Chakraborty, R., Fuerst, P.A. (1976). Proc. Nat. Acad. Sci. **73**, 4164–4168

Nei, M., Fuerst, P.A., Chakraborty, R. (1978). Proc. Nat. Acad. Sci. **75**, 3359–3362

Nemer, M., Dubroff, C.M., Graham, M. (1975). Cell **6**, 171–178

Nevins, J., Darnell, J.E. (1978). J. Virology **25**, 811–825

Nikolaev, N., Silengo, L., Schlessinger, D. (1973). Proc. Nat. Acad. Sci. **70**, 3361–3365

Pearson, E.S., Hartley, H.O. (1966, 1972). Biometrika tables for statisticians, Vol. 1 and 2., Cambridge: Cambridge University Press

Penman, S. (1966). J. Mol. Biol. **17**, 117–130

Penman, S., Scherrer, K., Becker, Y., Darnell, J.E. (1963). Proc. Nat. Acad. Sci. **49**, 654–662

Perry, R.P. (1963). Exp. Cell Research **29**, 400–406

Perry, R.P., Latorre, J., Kelly, D.E., Greenberg, J.A. (1972). Biochim. Biophys. Acta **262**, 220–226

Peterson, J.L., McConkey, L. (1976). J. Biol. Chem. **251**, 548–554

Polasa, H., Green, M. (1967) Virology 31, 565–567

Preston, F.W. (1948). Ecology **29**, 254–283

Proudfoot, N.J. (1977). Cell **10**, 559–570

Proudfoot, N.J., Gillam, S., Smith, M., Longley, J.I. (1977). Cell **11**, 807–818

Puckett, L., Darnell, J.E. (1976). J. Cell Physiol. **90**, 521–534

Reddy, V.B., Thimmappaya, B., Dhar, R., Subramanian, K.N., Zain, B.S., Pan, J., Ghosh, P.K., Celma, M.L., Weissman, S.M. (1978). Science **200**, 494–502

Reeck, G. (1976). In: CRC Handbook of biochemistry and molecular biology: Proteins. Fasman, G.D., ed., Vol. 3, pp. 504–519, Cleveland: CRC Press

Sanger, F., Dir, G.M., Barrell, B.G., Brown, B.L., Coulson, H.R., Fiddes, J.C., Hutchinson, C.V., Slocombe, P.M., Smith, M. (1976). Nature **265**, 687–698

Sawicki, S., Jelinek, W., Darnell, J.E. (1977), J. Mol. Biol. **113**, 219–239

Spradling, A., Hui, H., Penman, S. (1974). Cell 4, 131–137

Strauss, J.H., Kelly, R.B., Sinsheimer, R.I. (1968). Biopolymers **6**, 793–807

Sueoka, N. (1961). Proc. Nat. Acad. Sci. **47**, 1141–1149

U.S. Department of Commerce, Office of Business Economics. (1952). Income distribution in the United States, Washington, D.C.: US Govt. Printing Office

Vallee, B.L., Wacker, W.E.C. (1976). In: CRC Handbook of biochemistry and molecular biology: Proteins. Fasman, G.D., ed., Vol. 3, pp. 278–292, Cleveland: CRC Press

Villa-Komaroff, C., Guttman, N., Baltimore, D., Lodish, H.F. (1975). Proc. Nat. Acad. Sci. **72**, 4157–4161

Williams, C.B. (1937). Ann. Appl. Biol. **24**, 404–414

Yuan, P.T. (1933). Ann. Math. Statistics **6**, 20–34

Received May 25, 1979; Revised October 24, 1979