The Evolution of Proteins from Random Amino Acid Sequences: II. Evidence from the Statistical Distributions of the Lengths of Modern Protein Sequences

Stephen H. White

Department of Physiology and Biophysics, University of California, Irvine, CA 92717, USA

Received: 24 June 1992 / Revised: 24 May 1993

Abstract. This paper continues an examination of the hypothesis that modern proteins evolved from random heteropeptide sequences. In support of the hypothesis, White and Jacobs (1993, J Mol Evol 36:79–95) have shown that any sequence chosen randomly from a large collection of nonhomologous proteins has a 90% or better chance of having a lengthwise distribution of amino acids that is indistinguishable from the random expectation regardless of amino acid type. The goal of the present study was to investigate the possibility that the random-origin hypothesis could explain the lengths of modern protein sequences without invoking specific mechanisms such as gene duplication or exon splicing. The sets of sequences examined were taken from the 1989 PIR database and consisted of 1,792 “superfamily” proteins selected to have little sequence identity, 623 E. coli sequences, and 398 human sequences. The length distributions of the proteins could be described with high significance by either of two closely related probability density functions: The gamma distribution with parameter 2 or the distribution for the sum of two exponential random independent variables. A simple theory for the distributions was developed which assumes that (1) protoprotein sequences had exponentially distributed random independent lengths, (2) the length dependence of protein stability determined which of these protoproteins could fold into compact primitive proteins and thereby attain the potential for biochemical activity, (3) the useful protein sequences were preserved by the primitive genome, and (4) the resulting distribution of sequence lengths is reflected by modern proteins. The theory successfully predicts the two observed distributions which can be distinguished by the functional form of the dependence of protein stability on length.

The theory leads to three interesting conclusions. First, it predicts that a tetra-nucleotide was the signal for primitive translation termination. This prediction is entirely consistent with the observations of Brown et al. (1990a,b, Nucleic Acids Res 18:2079–2086 and 18:6339–6345) which show that tetra-nucleotides (stop codon plus following nucleotide) are the actual signals for termination of translation in both prokaryotes and eukaryotes. Second, the strong dependence of statistical length distributions on sequence-termination signaling codes implies that the evolution of stop codons and translation-termination processes was as important as gene splicing in early evolution. Third, because the theory is based upon a simple no-exon stochastic model, it provides a plausible alternative to a limited universe of exons from which all proteins evolved by gene duplication and exon splicing (Dorit et al. 1990, Science 250:1377–1382).

Key words: Protein folding — Protein sequence lengths — Protein evolution — Gene structure — Ex-
ons — Introns — Exon splicing — Limited universe of exons — Translation termination

Introduction

Two hypotheses have been advanced to explain the origin of modern proteins. The more conventional one is the starter-set hypothesis, which proposes that modern proteins originated from a small set of starter sequences with lengths of 30–50 amino acids which were expanded by gene duplication and modification (Eck and Dayhoff 1966; McLachlan 1972; Gilbert 1978; Doolittle 1978; Darnell 1978; Doolittle 1979; Blake 1983; Holland and Blake 1990; Dorit et al. 1990; Patthy 1991; Dorit and Gilbert 1991). The most important feature of this hypothesis is the assumption that gene or exon duplication was essential for producing modern sequences which have an average length of about 300 amino acids. In contrast, the random-origin hypothesis (White and Jacobs 1990; Lau and Dill 1990; Shakhnovich and Gutin 1990) invokes no specific mechanisms beyond the presumption of a primitive system for the coding, translation, and synthesis of protein sequences. Given such a system, White and Jacobs (1990) suggested that there may have been a protein synthetic "big bang" which resulted in a very large number of random heteropeptide "protosequences." Because the ability to fold into compact structures is not particularly restrictive (Chan and Dill 1990; Lau and Dill 1990; Shakhnovich and Gutin 1989, 1990), a significant fraction of these sequences could have folded into compact structures to form a protogroup of proteins from which the primitive proteins were selected as a result of their ability to catalyze some advantageous chemical reaction or serve a useful structural role.

The random-origin hypothesis has two corollaries. First, the amino acid sequences of the primitive proteins would be indistinguishable mathematically from sequences produced by a random independent process. Second, the lengths of the protosequences would be determined solely by nucleotide "begin" and "end" signals that delimited the primitive genes or their transcribed messages and that were distributed randomly along nucleotide sequences. White and Jacobs (1993) have shown that the modern protein sequences conform well to the expectations of the first corollary. The second corollary, and more crucial one, is examined in the present paper.

The lengths of protein sequences are important in the evolution of proteins because protein stability is determined in part by length (Dill 1985; White 1992). Indeed, the recognition of this fact is the major impetus for invoking gene or exon duplication in the starter-set hypothesis (McLachlan 1972; Blake 1983). Therefore, the statistical distributions of the lengths of proteins should provide useful information about the organization of the genome and its relation to the stability and evolution of proteins. If these distributions can be explained by a simple stochastic model without assuming gene duplication or exon splicing, then the random-origin hypothesis would be supported.

In pursuit of this goal, I examined the length distributions of three sets of protein sequences drawn from the Protein Identification Resource (PIR) database and the set of exons recently described by Dorit et al. (1990). The protein sets consisted of 1,792 "superfamily" sequences selected to have little sequence identity, 623 E. coli sequences, and 398 human sequences. The superfamily set was chosen to serve as a broad representative example of proteins while the E. coli and human sets were chosen to compare prokaryotic and eukaryotic proteins. The analysis reveals that all three protein sets have similar statistical distributions that are described accurately by the distribution for the sum of two exponential distributions or by the closely related gamma distribution (parameter 2). The exon length distribution cannot be described by either of these functions or closely related ones.

The statistical distribution of the sizes of proteins has been investigated by other workers. Nei et al. (1976) examined 119 proteins and found, interestingly, that the density of molecular weights could be accurately fitted by a type of gamma distribution although no rationale was given for this choice. Sommers and Cohen (1980) examined the size distribution of about 500 protein subunits and fitted them as log-normal distributions. Gamma distributions can frequently be fit by log-normal distributions because both functions have similar asymmetries. However, when the protein length data of the present paper are fit with log-normal distributions, the density of points is highly non-uniform. Sommers and Cohen argued that the log-normal was appropriate because it describes distributions that arise from the product of many random independent events. Considering that proteins increase in length by addition, the log-normal distribution appears to be inappropriate. Gamma distributions, on the other hand, result from the stochastic addition of random intervals (Ross 1989) and are therefore more reasonable a priori.

A very simple theory for the observed gamma-type distributions was derived from the following ideas. As shown in subsequent paragraphs, a natural consequence of the random-origin hypothesis is that the lengths of the protoprotein sequences would have been exponentially distributed regardless of whether they formed directly from amino acid polymerization or indirectly through a transcription/translation process. This statistical distribution places a strong upper limit on the lengths of protoprotein sequences. Another consequence of the hy-
hypothesis is that the production of protoproteins (i.e., random polypeptides) was independent of their ability to fold into the compact structures necessary for the development of biochemical activity. Dill (1985) has shown from first principles that protein stability depends on sequence length and that proteins shorter than about 70 amino acids are much less likely to be stable than longer ones. White (1992) has provided direct support for this conclusion by means of statistical studies of the length dependence of the frequencies of amino acids, particularly cysteine. Thus, two independent length effects would have been important in the primordial soup; one would determine the length distribution of the protoproteins while the other placed constraints on the folding of the protoproteins into compact structures. The statistical distribution of the sequence lengths of the resulting primitive proteins would thus have been determined by the interplay of these two effects. The theory presented combines these two independent processes and predicts the shape of the length distribution of the primitive proteins which can be fit with high significance to the data for modern proteins. Because the theory does not require assumptions about gene duplication or exon splicing to explain the observed distributions, it provides, in my opinion, strong support for the random-origin hypothesis.

Theory

Begin by considering the primordial genome as a random sequence of nucleic acids such as one resulting from a polymer condensation reaction in a well-mixed solution of nucleic acids. I will informally refer to such a sequence as a "random sequence" but the use of that term should be explicitly interpreted as a sequence of random independent events (nucleotides). Assume that certain combinations of nucleotides represent begin-signal for protein sequences while other combinations represent end-signal and thus define protogenes. To avoid the unknowable details of primordial transcription and translation, I assume that, on average, there is a linear relationship between the lengths of the protogenes and the lengths of the resulting protoproteins. For a random sequence, the begin- and end-signal will have a uniform random distribution so that the intervals between signals will have an exponential distribution (Ross 1989). This has been shown directly by Senapathy (1986) through simulations using random DNA sequences. Importantly, he has also shown that the distribution of reading frame lengths in natural DNA sequences is exponential. In this scenario of primordial protein synthesis, the random protein sequences are produced independent of their ability to fold. Those that do fold have the potential for biochemical activity and I assume that the "active" primitive proteins gave their primitive genes a selective advantage. If such genes are assumed to have survived at the expense of the "nonproductive" genes, the distribution of the lengths of the surviving genes and their proteins would eventually achieve a stable distribution that would reflect the combined effects of the starting distribution of gene/protein lengths and the constraints on protein folding related to sequence length. Because no assumptions are made with respect to exon splicing or other specific processes, I will refer to the model as the "no-exon" model.

Assume that the gene begin-signals occur with a frequency $\sigma \ll 1$ and end-signals with a frequency $\epsilon \ll 1$. The exponential distribution is "memoryless," so the probability of encountering a signal is independent of position. This means that one may choose an arbitrary position along the chain as a starting point for measuring the interval to a signal. It is convenient to choose a begin-signal. Further assume that the occurrence of another begin-signal prior to the first occurrence of an end-signal has no effect on the process. The intervals $Z$ between begin-signals and end-signals are exponentially distributed random variables so that the probability of a sequence of length $Z$ is given by

$$p_Z(z) = \epsilon \cdot \exp(-\epsilon z)$$  \hspace{1cm} (1)

If the probability $p_i(z)$ of a protein folding into a compact structure is dependent on $p_Z(z)$, the probability density function for folded proteins of length $Z$ is

$$P_Z(z) = p_i(z) \cdot p_Z(z)$$  \hspace{1cm} (2)

so the probability of a protein of length $Z$, having a length between $z$ and $z + dz$ is given by $P_Z(z)dz$. The probability function $p_i(z)$ is not known. However, the calculations of Dill (1985) suggest that maximum stability occurs at several hundred amino acids and declines for longer or shorter sequences. However, the formation of domains can extend the stability range and would tend to cause stability to be less dependent on length for long sequences. For computational purposes, I assume that $p_i(z)$ is a monotonically increasing function of a simple type and ignore the possibility of it declining for lengths greater than some particular value. This assumption is reasonable if the probability $P_Z(z)$ of a long message declines faster than the probability $p_i(z)$ of folding. I consider two forms for $p_i(z)$. One form is a simple linear function (eq. 3) and the other a simple exponential function (eq. 4):

$$p_{i1}(z) = k_1z$$  \hspace{1cm} (3)

$$p_{i2}(z) = k_2[1 - \exp(-\phi z)]$$  \hspace{1cm} (4)
where \( k_1 (\ll \epsilon) \) and \( k_2 \) are constants and \( \phi \) a parameter describing the strength of the dependence of the probability on \( z \). By substituting eq. 3 or 4 into eq. 2 and normalizing over the interval \( 0 \leq \epsilon \leq \infty \) so that \( \int_{-\infty}^{\infty} p_2(z)dz = 1 \) for the chains which fold, it is easy to show that the two forms of \( p_2(z) \) are

\[
p_2(z) = z\epsilon^2 \cdot \exp(-\epsilon z) \tag{5a}
\]

\[
p_2(z) = [((\phi + \epsilon)\epsilon/\phi][\exp(-\epsilon z) - \exp(-\{\phi + \epsilon\}z)] \tag{5b}
\]

The second equation (5b) is plotted in Fig. 1A (solid line) using parameters determined for the superfAMILY set of proteins. \( p_2(z) \) from eq. 1 and \( p_2(z) \) from eq. 4 are also plotted in Fig. 1A to show how the probability densities combine to affect the shape of the protein length-distribution curve.

I show below that either eq. 5a or 5b will generally fit the experimental data with a reduced \( \chi^2 = 1 \). To decide which function is the better one in a given case, a sensitive statistical test is required. Such a test, developed by Norbert Schumacher, is described in the appendix. It is based upon the fact that the probability density functions contained within eqs. 5a and 5b are, respectively, the gamma distribution (parameter 2) and the distribution for the sum of two exponentially distributed independent random variables. The equations can be recast into the general form appropriate for these problems by letting \( \epsilon = \alpha \) and \( \epsilon + \phi = \beta \). Then, in explicit statistical terms, if \( Z \) is the sum of two exponentially distributed independent random variables \( X \) and \( Y \) with parameters \( \alpha \) and \( \beta \), the probability density function \( f_Z(z) \) has two forms depending upon the relative values of the parameters (I ranpour and Chacon 1991). For \( z < 0 \), if \( \beta = \alpha \) then

\[
f_Z(z) = \alpha^2 \cdot z \cdot \exp(-\alpha z) \tag{6a}
\]

whereas if \( \alpha \neq \beta \)

\[
f_Z(z) = [\alpha\beta/(\beta - \alpha)][\exp(-\alpha z) - \exp(-\beta z)] \tag{6b}
\]

For \( z < 0 \), \( f_Z(z) = 0 \) in both cases. Norbert Schumacher’s appendix gives methods for estimating \( \alpha \) and \( \beta \) and for testing the null hypothesis \( H_0 \) that \( \alpha = \beta \) using a test statistic \( T \). \( H_0 \) is rejected in favor of the alternate \( H_1 : \alpha \neq \beta \) if \( T > 0.89 \) with a significance of \( P(T > T_0) < 0.01 \).

The theory just presented is based upon the assumption that nucleotide sequences preceded polypeptide sequences, but this is not necessary. One can arrive at an equation similar to eq. 1 without invoking nucleic acid participation by considering the formation of a random polypeptide from a condensation reaction in a solution of amino acids. The distributions of the lengths of the resulting polymers have been discussed by Flory (1953). He showed that the probability of obtaining a polymer comprised of \( x \) monomers is \((1 - p)^{x-1}\) where \( p < 1 \) is the extent of reaction defined as the probability of a covalent bond forming between two monomers. This equation may be transformed to one similar to eq. 1 by suitably defining constants, letting \( p = e^K \), and noting that \( K < 0 \) when \( p < 1 \). Thus, the distribution of polymer lengths in a random condensation will also be exponentially distributed. Flory’s analysis applies to any linear polymer and therefore suggests that the primordial nucleotide sequences should have had exponentially distributed lengths as well. Thus, whether one begins with random polypeptide or random nucleotide sequences, eqs. 5 and 6 will result providing that the
ability to fold into a compact structure is independent of the process that produced the starting sequence.

Methods

Three sets of protein sequences were drawn from release 21 of the 1989 International Protein Identification Resource (PIR) Sequence Database produced by the National Biomedical Research Foundation (Washington, D.C.). (See Barker et al., 1991.) The first set is the Superfamily Set described by White and Jacobs (1993), which consists of 1,792 sequences with little sequence identity which satisfy certain statistical criteria established to validate the use of the run test for examining the distributions of amino acids along the length of protein sequences. Basically, one sequence was drawn from each PIR superfAMILY and those used were required to be complete, unambiguous, and to have at least 21 residues. The present superfAMILY set contains slightly more proteins than the 1,789 reported by White and Jacobs (1993) because some minor constraints not relevant to the present work were removed. Because I was interested in comparing the length distributions of prokaryotic and eukaryotic sequences, all of the E. coli and human proteins satisfying the same statistical criteria as the Superfamily Set were also drawn from the complete database. I thus obtained the Human Set consisting of 398 sequences and the E. coli Set consisting of 623 sequences. The sequences were extracted from the database using programs written in the C language and run on a Sun 4/110 computer (Sun Microsystems, Sunnyvale, CA). The set of exon lengths were extracted from Fig. 1 of the paper by Dool it et al. (1990) concerning with the hypothesis that all modern proteins originated from a limited number of exons. These lengths comprise the Exon Set.

A vector of lengths was constructed for each set of proteins and prepared for analysis by sorting the lengths into bins of width $dz = 10$ amino acids to produce a number distribution $n(z)$ of sequence lengths. Probability density functions were fitted to the data by $x^2$ minimization using an 80386 microcomputer running the program FitAll (v. 5.01, MTR Software, Toronto, Ontario). The fits were performed using the equation $n(z) = f(z) \cdot N \cdot dz$ where $N$ is fixed at the number of sequences in the set and $f(z)$ is given by eq. 6. Because of the shape of the distribution of exon lengths, exon was fitted by letting $N$ be a free parameter. The test statistic $T$ described in the appendix was calculated by means of Mathcad (v. 3.0, Mathsoft, Cambridge, MA) for each set in order to test $H_0: \alpha = \beta$.

Results

Length Distributions of Protein Sequences

The number densities $n(z)$ of the lengths and the best-fit curves (solid lines) for the three protein sets are shown graphically in Fig. 2, panels A–C; the parameters for the fits are summarized in Table 1. Also shown in Table 1 are the mean and mode lengths of the various sets. All of the protein data sets could be fit with the probability density functions of eq. 6 with a reduced $x^2$ of 1.02 or less with high significance ranging from $P(x^2 > x^2_0) = 0.394$ for E. coli to 0.9998 for superfAMILY (Table 1). The protein sets thus adhere closely to theory.

The data of Table 1 for Superfamily, E. coli, and Human show that either eq. 6a or 6b describes the data about equally well in terms of the values of reduced $x^2$. However, the observed values $T_0$ of the test statistic $T$ (appendix) demonstrate for human and superfAMILY that eq. 6b is the correct choice because the T values of 1.93 and 6.50 are unequivocally greater than the value of 0.89 established for a significance $P(T > T_0) \leq 0.01$. The T value of $-0.04$ for E. coli demonstrates unequivocally that eq. 6a is the best description of those data. It should be noted that although the differences in the reduced $x^2$ values for eqs. 6a and 6b are small for the three sets of proteins excluding exon, the differences are in directions consistent with the result of the T statistic. Recalling that eq. 6a resulted from a linear probability-of-folding function and 6b from an exponential function, these results indicate that the probability of folding increases more rapidly with length for the human and superfAMILY sets than for the E. coli set. This is also indicated by the mode values (Table 1) of 126 and 115 for Human and Superfamily compared to 151 for E. coli.

The greater rate of increase in folding probability for
superfamily and human appears to be due to disulfide bonds. The frequencies of cysteine are significantly higher in human (0.025) and superfamily (0.017) than in E. coli (0.011) (unpublished observation). More significantly, a careful examination of the frequency of cysteine in a superfamily set of proteins from the 1990 PIR database shows (1) that Cys frequency increases drastically for short protein sequences compared to long ones and (2) that the increase corresponds accurately to the increase expected for an average of two Cys per sequence (White 1992). A similar analysis of amino acid frequencies for the protein sets used in the present study reveals a similar behavior of Cys for superfamily and human. Importantly, E. coli shows little, if any, dependence of Cys frequency on length because, apparently, the reducing environment inside E. coli makes disulfide bonds less useful as a means of achieving stability. It seems clear that E. coli must rely on noncovalent effects such as increases in sequence length to achieve stability (Table 1). Increases in Arg and Lys frequencies also apparently contribute to the stability of small proteins in general and small E. coli proteins in particular (White 1992).

Taken together, the results for the three sets of proteins show a great deal of internal consistency. Because the fits of the data to the distributions also have high statistical significance, one may conclude that the sampling of the PIR database is robust. That is, the shapes of the distributions are unlikely to have resulted from sampling errors and are also unlikely to change significantly as the sizes of samples increase. A 1990 superfamily set of 2,275 sequences is accurately fit by the distribution describing the 1989 superfamily set (White 1992).

Although I chose the probability of folding functions $p_f(z)$ shown in eqs. 3 and 4 largely for convenience, the excellent fits of the data to theory suggest that they may be reasonably accurate. This conclusion is supported by the fact that higher-order gamma-distribution functions cannot be fit to the data satisfactorily, which I interpret to mean that $p_f(z)$ functions that depend upon powers of $z$ greater than 1 will not be satisfactory. Dill's (1985) analysis of the dependence of the stability of folded proteins on length suggests that $p_f(z)$ should pass through a maximum. This is probably correct but, as suggested in Theory earlier, this will not have a major effect if the length probability function $p_L(z)$ of eq. 1 declines more rapidly than the increase in $p_f(z)$. The satisfactory fit of the theory of the data suggests that this is the case. Therefore, it appears that the rising phase of the length distribution curve is dominated by length restrictions on folding and the falling phase by the decreasing probability of long protein/gene
lengths. This is summarized graphically in Fig. 1A
where $P_s(z) \cdot N$, $p_s(z) \cdot N$, and $p_{r2}(z)$
have been plotted using the parameters from Table 1
for the superfamily
two-exponential fit. [The probability density $p_s(z)$
has been scaled arbitrarily to make it visible on the graph
which displays number density rather than probability.]

**Length Distribution of Exon Sequences**

The number density $n(z)$ of the exon lengths and the
best-fit gamma distribution (parameter 2, dashed line)
are shown in Fig. 2D. Although the length distribution
of Exon appears at first glance to be similar to those of
the protein sets, it is actually quite different and does not
adhere to theory. The reduced $\chi^2$ values of 6.46
and 7.11 for the gamma(2,0) and two-exponential fits (Table 1)
leads to the rejection of the hypothesis that the theory
fits the data with a significance $P(\chi^2 > \chi^2_o) < 10^{-6}$.
The exon data could be fit somewhat better by gamma distributions
with parameter 3 or 4 but for either case, the
reduced $\chi^2$ was approximately 3, meaning that these
higher-order curves do not describe the data with a signification $P(\chi^2 > \chi^2_o) \ll 0.01$. The best-fit gamma distribution
with parameter 4 is shown by the solid line in
Fig. 2D.

An important aspect of Exon is that its mean and
mode lengths of 51 and 35, respectively, differ considerably
from the corresponding parameters of the protein
sets. Only a small proportion of the set (Fig. 2D) have
lengths of 80 or more amino acids and the maximum
length reported is less than 140. Dill’s theory (1985) argues
against proteins with such short sequences being
generally stable without disulfide bonds and/or ligands.
Thus, if exons represent structural or functional
domains of mature proteins (Rossman 1990; Holland and
Blake 1990), it seems unlikely that the polypeptides
coded for by them would generally have had an independent
existence as proteins at an earlier stage of evolution unless they were rich in cysteine or had a strong
ability to bind ligands (White 1992).

**Discussion**

**Implications for the Random-Origin and
Starter-Set Hypotheses**

The results of the analysis presented above show that the
length distributions observed for modern proteins are
accurately described by the theory. This provides support
for the random-origin hypothesis. However, an important
question is whether or not a similar result can be obtained
through some version of the starter-set hypothesis.
Although detailed examinations of alternatives to the
random-origin hypothesis are hindered by our profound
lack of knowledge about the nature of gene duplication
and exon splicing in the early biosphere, there is one
particularly simple and obvious possibility that can be
discussed that I refer to as the “stochastic-exon model.”
Assume that the primitive proteins were assembled
by randomly splicing together short peptide sequences that
might, for example, have been coded for by primitive exons.
Further assume that these primitive starter sequences
have exponentially distributed lengths as would be
expected either by a process of random copolymerization
or by transcription/translation with uniform randomly
distributed begin- and end-signals. Two results follow immediately from this scenario. First, the length
distribution of the primitive exons would be exponential.
(See Theory section.) Second, the length distribution
of the resulting proteins would be some form of gamma distribution because the proteins are assembled
by an additive stochastic process equivalent to a multi-event waiting time problem. (See, for example, Ross,
1989.) It is easy to show that if proteins were assembled exclusively from 2, 3, 4, etc., exon-peptides, the resulting
distributions of protein lengths would be gamma distributions with parameters 2, 3, 4, etc. Unless the predominant
number of exons per gene were exactly two, the density of protein sequence lengths would therefore have been described by gamma-type distributions of order higher than two. This simple analysis reveals two important issues that must be considered in the development of more elaborate stochastic starter-set models:
The length distribution of exons and the number density of exons in primitive genes. It is obvious that the modern eukaryotic protein length distribution must be derivable from the modern exon length and number densities. The challenge for the starter-set hypothesis is to explain the origin of these densities by means of a stochastic model.

**The Determinants of the Maximum Lengths of
Modern Proteins**

The conclusion that the modern protein length distributions can be explained without invoking exon splicing
differs from the conclusion of Senapathy (1986). Based upon extensive statistical analyses of natural and
simulated DNA sequences, he showed that the observed distribution for DNA reading-frame lengths (RFL) is
accurately exponential and the same as for the intervals between three-codon amino acid codons. As noted earlier,
this observation supports the notion that “signals” are uniformly randomly distributed along natural nucleic
acid sequences as assumed in the random-origin hypothesis. Although the length distribution of the exon set
is not exponential, the general size range of exons is consistent with Senapathy’s observation that the maximum
RFL for eukaryotic DNA is 200 codons. Armed with this observation, Senapathy argued (1986, 1988) in

```
favor of "introns-early" by suggesting that there was considerable evolutionary pressure for intron splice-out to remove stop codons in order to increase the length distribution of proteins.

Senapathy's hypothesis hinges upon the assumption that three trinucleotide stop codons were used in the primitive genetic code. If this is correct, then one can easily calculate the expected distribution of intervals between stop codons from the probability function of eq. 1 by setting \( e \) equal to the probability of occurrence of stop codons in a random sequence which is \( 3 \cdot (1/4)^3 = 0.047 \). The distribution for stop signals encoded by three trinucleotides (3NT) is plotted in Fig. 1B as a dashed (- - -) curve. Note that, consistent with Senapathy's findings, the probability of lengths greater than 200 codons is small. But, there is no reason to believe that there were three 3NT stop codons in the primitive genome. If there had only been only one 3NT codon, as is the case for Tetrahymena thermophila (Hanyu et al. 1986), the probability of occurrence would have been \( e = (1/4)^3 = 0.016 \), which significantly increases the maximum interval as shown in Fig. 1B by the dot-dash (---) curve. But, one might go even further and assert that the primitive stop codons were not encoded by trinucleotides. If the primordial signals consisted of four, five, or six nucleotides the frequencies would be \( (1/4)^4 = 0.0039 \), \( (1/4)^5 = 0.00098 \), and \( (1/4)^6 = 0.00024 \). As can be seen from the length distribution for a single tetra-nucleotide (4NT) signal plotted as a solid line in Fig. 1B, this would drastically increase the maximum protein length. Interestingly, based on the parameters of Table 1 for the two-exponential fit to the Superfamily data, \( e = 0.0048 \), which is close to the probability for a 4NT signal. This parameter was used in calculating the protein length distribution in Fig. 1A (solid line), which can be seen to have a "tail" with an extent similar to that for the 4NT signal in Fig. 1B (solid line).

The observation that the length limit of modern proteins can be explained by a tetra-nucleotide stop signal is quite remarkable in light of studies which show that the efficiency of the translation of stop codons by suppressor tRNA is context dependent. (See, e.g., Bossi and Roth, 1980.) Recently, Brown et al. (1990a,b) have done extensive statistical studies of these context effects and have shown that the termination of translation is strongly affected in both prokaryotes and eukaryotes by the nucleotide immediately following the stop codon. So significant and widespread is the effect that Brown et al. (1990b) proposed that "the stop codon and the nucleotide following it comprise a tetra-nucleotide stop signal." These context effects apparently result from the effect of the tetra-nucleotide signals on the binding of release factors. (See review by Tate and Brown, 1992.) The significance of the observation that tetra-nucleotides signal translation termination in both prokaryotes and eukaryotes should not be overlooked. This means that this signal predates the prokaryote/eukaryote divergence and must therefore have been operative in the common ancestor.

These considerations tie the evolution of the lengths of protein sequences to the evolution of stop and nonsense codons and the translation-termination processes. This connection is extremely interesting in the context of the work of Jukes and his colleagues (Jukes 1982; Jukes et al. 1987; Osawa and Jukes 1988), who have studied the evolution of genetic codes using the existing differences between organelle and cellular codes. Changes in the code for stop codons are an essential part of Jukes's (1982) scheme of codon evolution as a means of expanding and refining the code without the creation of disruptive ambiguities. These studies emphasize that the genetic code has an evolutionary history and that changes in stop codes are of central importance to it. There is thus no reason to assume that three tri-nucleotide stop codons have been an invariant feature of genomes.

The apparently correct prediction of a tetra-nucleotide end-signal by the theory presented in this paper provides encouraging support for the random-origin hypothesis and suggests that the evolution of stop codons and translation-termination processes may have been of greater importance to the early evolution of proteins than gene duplication or exon shuffling. Nevertheless, given that these two processes are known to be important in the evolution of eukaryotes, there should be some evidence for them from the length distributions. A close examination of Fig. 3 reveals that the major difference between E. coli and human is the somewhat longer tail of the human protein length distribution. This lengthening may have occurred as result of gene duplication and exon shuffling as a later process of evolution.

**Introns Early or Introns Late?**

There is strong evidence that many proteins have originated from intron-mediated recombination of exons that code for functional and structural domains of proteins as proposed by Gilbert (1978). (See reviews by Holland and Blake [1990] and W.F. Doolittle [1990].) A major unanswered question is whether exons appeared early (Doolittle 1978; Blake 1983) or late (Cavalier-Smith 1985) in evolution relative to the prokaryote/eukaryote divergence. Gilbert and his colleagues (Dorit et al. 1990; Dorit and Gilbert 1991) favor the view that introns are an ancient feature of genomes that may have been present before the first branchings of life took place. Furthermore, they have suggested that all modern proteins originated by recombination of exons...
from a "limited universe of exons" that could have numbered as few as 1,000–7,000 exons (Dorit et al. 1990). (But see criticisms by R.F. Doolittle [1991] and Patthy [1991].)

Irrespective of the concept of a limited universe of exons, it is interesting that no qualitative differences in the distributions of the lengths of human and E. coli proteins were found despite the major differences in gene organization and mRNA synthesis that exist between eukaryotes and prokaryotes (Fig. 3). The quantitative differences that do exist are easily accounted for within the framework of the no-exon model by simple protein folding requirements. The only possible conclusion from this result is that the length information stored in the genomes of eukaryotes and prokaryotes is fundamentally independent of the existence of intervening sequences. If one accepts the random-origin hypothesis and the no-exon model, as the close agreement between theory and data seem to demand, then one must conclude that the length information was embedded in gene structure very early in evolution and prior to the advent of introns. On the other hand, if one rejects the random-origin hypothesis in favor of the starter-set hypothesis, then one can argue that the similarity of the E. coli and human length distributions demonstrates that intron-mediated evolution has been present from the beginning. However, to adopt this position, one must contrive an intron-mediated theory that can reasonably explain the observed length distributions. Although I am not optimistic that this can be done, attempts should be made.

A careful and extensive reexamination of existing data (Naora and Deacon 1982; Blake 1983; Hawkins 1988; Traut 1988; Smith 1988) on the statistics of the lengths of genes, mRNA, exons, and introns may suggest models that are not presently obvious. However, an extra demand placed on such models is that they must predict the shape of the distribution of exon lengths as well as the lengths of the resultant proteins.

Given the simplicity of the random-origin hypothesis and no-exon model, I am inclined to believe that introns were not present from the beginning, which supports, perhaps, the transposon theory of Cavalier-Smith (1985). Because simplicity would have been highly advantageous in a feeble biosphere, the theory is particularly appealing. But, if introns were not present early, one must then explain how introns made their entrance without disrupting the length information stored in the genome of the common ancestor of prokaryotes and eukaryotes. Whatever the process was, it logically must have involved RNA splicing coupled with reverse transcription (Sharp 1985). One possibility is that the ancestor cells may have been subjected to a virus-like invasion whose legacy was introns. The lack of disruption of the information stored in the genome might then be explained by simple selection: Only those organisms with largely undisturbed protein length information survived. One can speculate that such an event could have marked the starting point for eukaryote/prokaryote divergence.

Implications for a Limited Universe of Exons

The results presented in this and a previous paper (White and Jacobs 1993) provide a coherent view of the early origin of modern protein sequences which I regard as strong support for the random-origin hypothesis (White and Jacobs 1990; Lau and Dill 1990; Shakhnovich and Gutin 1990). Several lines of evidence were of particular importance in arriving at that opinion. First, Lau and Dill (1990) and Shakhnovich and Gutin (1990) have shown through extensive simulations and analyses that random polypeptide sequences have a significant probability of folding into compact structures. This conclusion is supported by the fact that apparently 90% or more of known proteins have lengthwise distributions of amino acids that cannot be distinguished from the random expectation by means of the run and other tests (White and Jacobs 1990, 1993). Second, the shapes of the distributions of the run-test statistic for various combinations of amino acids in the modern proteins are close to N(0, 1) (i.e., normal with a mean = 0 and variance = 1) (White and Jacobs 1993). This shape would not be expected a priori unless the primitive proteins were produced by a single random process. Proteins that started as nonrandom sequences could subsequently become random because of random mutations. However, if all proteins followed such a course, one would not expect the distribution of the run-test statistic to have
any particular shape. Third, the shapes of the distributions of the lengths of modern proteins, independent of whether they are prokaryotic or eukaryotic, are predicted by a simple stochastic model independent of assumptions about gene duplication or exon splicing. Such processes can certainly cause increases in the lengths of protein sequences. But, if they were the principal means for increasing length, why should the distributions of the lengths of modern proteins adhere to the simple statistical function observed? And, why should the shapes of the prokaryotic and eukaryotic length distributions be so similar? Fourth, the simple no-exon stochastic model explains the upper limits on the lengths of protein sequences as due to the code for stop codons and predicts that the stop code is a tetra-nucleotide. This prediction is confirmed in modern sequences by the studies of Brown et al. (1990a,b) and is consistent with the broader idea that stop codons have played a key role in the evolution of genetic codes (Oswa and Jukes 1988). Importantly, tetra-nucleotide sequences are used to signal translation termination in both prokaryotes and eukaryotes. Because the signal is ancient, it provides the means of unifying the length distributions of E. coli and Human.

The random-origin hypothesis makes it unnecessary to assume that primitive proteins were assembled by gene duplication and exon shuffling from a set of small starter sequences in order to achieve diversity in the lengths and folding motifs observed in modern protein sequences. It must surely be true, however, that gene duplication and exon shuffling have been important in protein evolution (Gilbert 1978; Doolittle 1978; Blake 1983). The random-origin and starter-set hypotheses can be reconciled by replacing the smaller starter set of proteins with an “initial set” produced according to the random-origin hypothesis. This and the previous paper (White and Jacobs 1993) suggest that the repertoire of proteins in the initial set would have been large and diverse from the beginning. This would preclude the necessity for a small and limited “universe of exons” as proposed by Dorit et al. (1990).

Although Jacques Monod’s (1971) classic treatise on the evolution of life is generally found in book shops in the philosophy department, the work presented here supports his broad assessment of protein evolution as “randomness caught on the wing, preserved, reproduced, by machinery of invariance and thus converted into order, rule, necessity.”

Acknowledgments. I am grateful for the advice and criticisms of the “Wednesday Afternoon Group” consisting of Prof. Howard Tucker, Prof. Mark Finkelstein, Mr. Norbert Schumacher, and Mr. Les Vernon of the Mathematics Department. Dr. Jay Edelman provided useful comments on early versions of the manuscript. I thank Dr. Wes Hatfield for bringing to my attention the literature on codon context. The research was supported in part by the National Science Foundation (DMB-8807431) and the National Institutes of Health (GM-46823).

References

Brown CM, Stockwell PA, Trotman CNA, Tate WP (1990b) Sequence analysis suggests that tetra-nucleotides signal the termination of protein synthesis in eukaryotes. Nucleic Acids Res 18:6339–6342
Doolittle WF (1978) Genes in pieces: were they ever together? Nature 272:581–582
Hanyu N, Kuchino Y, Nishimura S (1986) Dramatic events in ciliate evolution: alteration of UAA and UAG termination codons to glutamine codons due to anticodon mutations in two Tetrahymena iRNAs(Gin). EMBO J 5:1307–1311
Let Z be the sum of two exponentially distributed independent random variables X and Y with parameters \( \alpha \) and \( \beta \), respectively. The probability density \( f_Z(z) \) has two forms depending upon the relative values of \( \alpha \) and \( \beta \).

Case 1, \( \alpha \neq \beta \):
- For \( z \geq 0 \):
  \[
  f(Z; \alpha, \beta) = \frac{\alpha \beta}{\alpha + \beta} \exp(-z \alpha - \beta z) 
  \]
  \[ \text{for } z < 0 \quad f(Z; \alpha, \beta) = 0 \] (A1)

Case 2, \( \alpha = \beta \) [a gamma distribution to \( Z \leq x \)]:
- For \( z \geq 0 \):
  \[
  f(Z; \alpha, \beta) = \frac{\alpha^z \beta^z}{\Gamma(z+1)} \exp(-z \alpha) 
  \]
- For \( z < 0 \):
  \[ f(Z; \alpha, \beta) = 0 \] (A2)

I seek estimates \( \hat{\alpha} \) and \( \hat{\beta} \) for \( \alpha \) and \( \beta \), respectively, and a test of the null hypothesis \( H_0 \) that \( \alpha = \beta \).

It is easy to show that the expectation values for \( Z \) and \( Z^2 \) are:

\[
E(Z) = \frac{\alpha}{1 - \beta} 
\]
\[
E(Z^2) = 2 \left( \frac{1}{\alpha^2} + \frac{1}{\alpha - \beta} \right) 
\]

The estimates \( \hat{\alpha} \) and \( \hat{\beta} \) can be obtained in the following manner. Let \( \mu_1 = E(Z) \) and \( \mu_2 = E(Z^2) \). Then, from eqs. A3 and A4, one may write:

\[
1/\beta = \mu_1 - 1/\beta 
\]
\[
\mu_2 = 2(\mu_1 - 1/\beta)^2 + (1/\beta)(\mu_1 - 1/\beta) + 1/\beta^2 
\]

Eq. A6 can be rearranged to yield:

\[
1/\beta^2 - \mu_1(1/\beta) + (\mu_1^2 - \mu_2)/2 = 0 
\]

which can be solved by the quadratic formula for \( 1/\beta \). After inversion, one obtains:

\[
\beta = 2/(\mu_1 \pm \sqrt{(2\mu_2 + 3\mu_1^2)}) 
\]

By a similar argument,

\[
\alpha = 2/(\mu_1 \pm \sqrt{(2\mu_2 + 3\mu_1^2)}) 
\]

Thus, one may choose:

\[
\alpha = 2/(\mu_1 + \sqrt{(2\mu_2 + 3\mu_1^2)}) 
\]
\[
\beta = 2/(\mu_1 - \sqrt{(2\mu_2 + 3\mu_1^2)}) 
\]

The estimates are obtained from eqs. A10 and A11 by noting that \( n^{-1} \sum Z_i \) converges in probability to \( E(Z) \) and \( n^{-1} \sum Z_i^2 \) to \( E(Z^2) \) where the summations on \( i \) extend from \( i = 1 \) to \( i = n \). The estimators for \( \alpha \) and \( \beta \) may therefore be taken as:

\[
\alpha^* = 2/(n^{-1} \sum Z_i + \sqrt{(2(n^{-1} \sum Z_i^2) + 3(n^{-1} \sum Z_i^2))}) 
\]
\[
\beta^* = 2/(n^{-1} \sum Z_i - \sqrt{(2(n^{-1} \sum Z_i^2) + 3(n^{-1} \sum Z_i^2))}) 
\]

The preceding arguments indicate that under \( H_0 \), \( \alpha = \beta \), \( 2E(Z^2) = 3E(Z)^2 \) while under \( H_1 \), \( \alpha \neq \beta \) (the alternate hypothesis), \( 2E(Z^2) > 3E(Z)^2 \). Thus, under \( H_0 \), \( E(Z)^2/E(Z^2) = 3/2 \). It can be shown that

---

**Appendix: Estimates for the Parameters of the Distribution for the Sum of Two Exponentially Distributed Independent Random Variables**

**Norbert Schumacher**

Department of Mathematics, University of California, Irvine, CA 92717, USA
\[ \text{Var}(Z_i^2/(E(Z))^2) = 5.25. A suitable test statistic for the null hypothesis is therefore} \]

\[ T = \frac{\sum(Z_{ij}^2/(E(Z))^2) - \frac{3}{2} n}{\sqrt{(5.25n)}} \quad (A14) \]

which should have a normal distribution with mean 0 and variance 1 [the N(0,1) distribution] under the Central Limit Theorem. However, E(Z) is unknown. I therefore adopt \( n^{-1}Z \) as a surrogate and determine the actual distribution by simulation. The simulation for \( 1 \leq j \leq 3,600 \) taken on

\[ T_j = \frac{\sum(Z_{ij}^2/(n^{-1}\sum Z_{ij}^2) - \frac{3}{2} n}{\sqrt{(5.25n)}} \quad (A15) \]

for \( i = 1 \) to \( i = n = 1,000 \) shows that the distribution of \( T_j \) is \( N(0, 0.145) \) with \( \max(T_j; 1 \leq j \leq 3,600) = +1.49 \) and \( \min(T_j; 1 \leq j \leq 3,600) = -1.23 \).

Using eq. A15, \( H_0: \alpha = \beta \) is rejected in favor of \( H_0: \alpha \neq \beta \) if the observed \( T \), \( T_0 \), is too large. Because \( T_j \) is distributed \( N(0, 0.381^2) \), \( H_0: \alpha = \beta \) can be rejected with significance \( P(T > T_0) < 0.01 \) if \( T_0 > 0.89 \).
where \( k_1 (\ll e) \) and \( k_2 \) are constants and \( \phi \) a parameter describing the strength of the dependence of the probability on \( z \). By substituting eq. 3 or 4 into eq. 2 and normalizing over the interval \( 0 \lesssim z \lesssim \infty \) so that \( |P_0(z)dz| = 1 \) for the chains which fold, it is easy to show that the two forms of \( P_0(z) \) are

\[
P_0(z) = Ze^2 \cdot \exp(-ez) \tag{5a}
\]

\[
P_0(z) = [(\phi + e)/\phi]\{\exp(-ez) - \exp(-\phi ez)\} \tag{5b}
\]

The second equation (5b) is plotted in Fig. 1A (solid line) using parameters determined for the superfam-ily set of proteins. \( P_0(z) \) from eq. 1 and \( P_0(z) \) from eq. 4 are also plotted in Fig. 1A to show how the probability densities combine to affect the shape of the protein length-distribution curve.

I show below that either eq. 5a or 5b will generally fit the experimental data with a reduced \( \chi^2 = 1 \). To decide which function is the better one in a given case, a sensitive statistical test is required. Such a test, developed by Norbert Schumacher, is described in the appendix. It is based upon the fact that the probability density functions contained within eqs. 5a and 5b are, respectively, the gamma distribution (parameter 2) and the distribution for the sum of two exponentially distributed independent random variables. The equations can be recast into the general form appropriate for these problems by letting \( e = \alpha \) and \( \phi = \beta \). Then, in explicit statistical terms, if \( Z \) is the sum of two exponentially distributed independent random variables \( X \) and \( Y \) with parameters \( \alpha \) and \( \beta \), the probability density function \( f_2(z) \) has two forms depending upon the relative values of the parameters (I ranpour and Chacon 1991). For \( z \gtrsim 0 \), \( \beta \geq \alpha \) then

\[
f_2(z) = \alpha^2 \cdot z \cdot \exp(-\alpha z) \tag{6a}
\]

whereas if \( \alpha \neq b \)

\[
f_2(z) = [\alpha/\beta/\beta - \alpha]\{\exp(-\alpha z) - \exp(-\beta z)\} \tag{6b}
\]

For \( z < 0 \), \( f_2(z) = 0 \) in both cases. Norbert Schumacher's appendix gives methods for estimating \( \alpha \) and \( \beta \) and for testing the null hypothesis \( H_0 \) that \( \alpha = \beta \) using a test statistic \( T \). \( H_0 \) is rejected in favor of the alternate \( H_1 : \alpha \neq \beta \) if \( T > 0.89 \) with a significance of \( P(T > T_{0.01}) < 0.01 \).

The theory just presented is based upon the assumption that nucleotide sequences preceded polypeptide sequences, but this is not necessary. One can arrive at an equation similar to eq. 1 without invoking nucleic acid participation by considering the formation of a random polypeptide from a condensation reaction in a solution of amino acids. The distributions of the lengths of the resulting polymers have been discussed by Flory (1953). He showed that the probability of obtaining a polymer comprised of \( x \) monomers is \( (1 - p)p^{x-1} \) when \( p < 1 \) is the extent of reaction defined as the probability of a covalent bond forming between two monomers. This equation may be transformed to one similar to eq. 1 by suitably defining constants, letting \( p = e^K \), and noting that \( K < 0 \) when \( p < 1 \). Thus, the distribution of polymer lengths in a random condensation will also be exponentially distributed. Flory's analysis applies to any linear polymer and therefore suggests that the primordial nucleotide sequences should have been exponentially distributed as well. Thus, whether one begins with random polypeptide or random nucleotide sequences, eqs. 5 and 6 will result providing that the