Origin of biological information: Inherent occurrence of intron-rich split genes, coding for complex extant proteins, within pre-biotic random genetic sequences

Periannan Senapathy*,1,2 Brajendra Kumar2, Chandan Kumar Singh2, Sudar Senapathy1, Bipin Balan2 and Raj Kuppuswami1

The origin of biological information is an unexplained phenomenon. Prior research in resolving the origin of proteins, based on the assumption that the first genes were contiguous prokaryotic sequences has not succeeded. Rather, it has been established that contiguous protein-coding genes do not exist in practically any amount of random genetic sequences. We found that complex eukaryotic proteins could be inherently encoded in split genes that could exist by chance within mere micrograms to milligrams of random DNA. Using protein amino acid sequence variability, codon degeneracy, and stringent exon-length restriction, we demonstrate that split genes for proteins of extant eukaryotes occur extensively in random genetic sequences. The results provide evidence that an abundance of split genes encoding advanced proteins in a small amount of prebiotic genetic material could have ignited the evolution of the eukaryotic genome.

Summary

This study presents evidence that split genes coding for highly complex proteins could have occurred indigenously in a small amount of pre-biotic random genetic sequences, possibly solving the enigma of the origin of biological complexity.

Introduction

While the origin of genes and proteins is still unknown, it has been traditionally assumed that a genome containing primitive, intronless, contiguous genes coding for simple proteins arose pre-biotically in a bacterial-like life form. Following its inception, this putative ancestor has been assumed to have evolved, in a tree-like linear branching pattern, into all of the complex genomes of the biota by random genetic mutation (*1-7*). This Linear Branching Evolution (LBE) model has been the prevailing paradigm of modern biology. In contrast, recent comparative genomics and proteomics studies have led to unexpected discoveries that challenge these assumptions. For example, studies examining the nature of domains, proteins, and protein fold superfamilies across phyla have shown that the earliest life-forms must have had sophisticated, eukaryote-like, proteins (*8-24*). In accordance, numerous complex proteins and domains that are widely distributed in nature have been found at the base of the evolutionary tree (*8-24*). Phylogenomic studies have also demonstrated the presence of a eukaryote, rather than a prokaryote, at the base of the evolutionary tree (*25-37*).

Contrary to prior beliefs that primitive split genes must have been present in the eukaryotic ancestor that became increasingly complex through evolution, interkingdom analyses have shown that the eukaryotic ancestor had a highly complex intron-rich genome akin to that of the modern eukaryotes (*38-39*). Furthermore, the genes of basal eukaryotes (e.g., trichoplax and sea anemone) were found to be as complex and intron-rich as those of higher animals (e.g., human) (*40-43*). Also, contrary to the earlier idea that the splicing process and the spliceosome must have been very simple in the eukaryotic ancestor, increasing in complexity through evolution, current evidence shows that the spliceosome of the earliest eukaryote must have been extremely complex and fully developed, with the same structure, function, and the >200 proteins and five non-coding RNAs present in modern higher eukaryotes (*44-47*). These discoveries demonstrate that the genomes and proteomes of the very first life forms may have been much more complex than originally hypothesized, and moreover may have been eukaryotic.

Consistent with these findings, emerging evidence concerning the character of several cellular components indicates that a eukaryote did not evolve from a prokaryote. Recent findings show that the RNA processing machinery within the eukaryotic cell nucleus, for example, is far too complex to have arisen by any endosymbiotic event from any prokaryote (*44-47*). Furthermore, the eukaryotic nucleus itself, with a unique membrane structure, integrated with the nuclear pore and the endoplasmic reticulum, lacks a clear homologue or precursor among prokaryotes (*48-52*). Several eukaryotic protein families

[1] Department of Human Genetics, Genome International Corporation, 8000 Excelsior Drive, Madison, WI 53717, United States of America
[2] Department of Bioinformatics, International Center for Advanced Genomics and Proteomics, Chennai, India

also lack a connection to any prokaryote. These entirely new proteins (and genes) could not have evolved even over 10^{35} years based on the empirically known rate of point mutation (10^{-9} to 10^{-6} per base per generation) and other mechanisms of genetic mutation (*53*), strengthening the idea that the eukaryote did not evolve from a prokaryote.

Prior attempts to solve the origin of biological information may have faltered due to the assumption that the first genes were contiguous (*2-7*). Here we employ an entirely different approach and explore whether split genes coding for complex proteins may have inherently occurred within prebiotic random genetic sequences. The analyses were conducted with split genes that encoded model proteins (containing known extant domains), complete extant multi-domain proteins, and extant proteins with sequence repetition. We incorporated amino acid variability (AAVAR) and codon degeneracy (CD) commensurate with that in extant genes and proteins. We searched for sequences encoding these proteins in split form with exon lengths similar to those found in extant intron-rich genes. Our findings demonstrate that complete split genes encoding complex proteins could have arisen within a minute amount of pre-biotic random DNA, explaining the origin of biological information and serving as the basis for the evolution of the very first genome.

Probability of split genes versus contiguous genes

An informational sequence has a considerably higher probability of occurring within a random sequence if it is split into short pieces than if it is contiguous. For example, the probability of "TO BE OR NOT TO BE" occurring as a single stretch in a random stream of English letters is $1/26^{13}$ (26 alphabetical letters with 13 characters in the phrase). The expected mean length (EML) of a random sequence that would, on average, contain one copy of this phrase is approximately the inverse of its probability:

$EML_{PHRASE} = EML_{TOBEORNOTTOBE} = 26^{13} = 10^{18}$ characters

When words can be separated by intervening random strings of characters, the EML for the phrase can be expressed as a sum of its component words. The probability of the occurrence of the split phrase in a random sequence increases significantly. Correspondingly, there is a substantial reduction in its EML. Thus the EML of the phrase "TO BE OR NOT TO BE" when split is:

$EML_{phrase} = EML_{TO} + EML_{BE} + EML_{OR} + EML_{NOT} + EML_{TO} + EML_{BE}$

$EML_{phrase} = 26^2 + 26^2 + 26^2 + 26^3 + 26^2 + 26^2 = 20{,}280$ characters

Applying this calculation to genetic material, the EML of the coding sequence of a gene reduces drastically if it is encoded in split form than if it is encoded in a contiguous sequence.

$EML_{intron\text{-}less\ gene} = EML_{continuous\ coding\ sequence}$

$EML_{split\ gene} = EML_{exon\ 1} + EML_{exon\ 2} + \ldots + EML_{exon\ n}$

The EML of a gene sequence coding for a 400 amino acid protein sequence without splitting (i.e., 1200 base contiguous coding sequence with no splits) would be exceptionally immense, 4^{1200} (10^{720}) bases, a length not encountered in nature (*2-7*). By contrast, the same sequence becomes probable within biologically plausible DNA lengths when it is split into short segments. We have termed this random origin of split biological information the random-sequence origin of split genes (ROSG) model (*54-57*).

Effect of CD and AAVAR on the probability of a gene sequence

Protein sequences are known to exhibit highly variable amino acid composition (up to 90%) (*53, 58*). That is, certain amino acids in a protein sequence can be substituted with one or more different amino acids without affecting the form and function of the protein. The genetic code is also redundant in that certain amino acids are encoded by multiple codons. The combination of AAVAR and CD can have a dramatic effect on the EML of its gene.

To illustrate, we used the λ repressor protein (Supplementary Information Figure 1) (*58*) and the known protein domains in the PFAM database (*59*) for which AAVAR had been previously documented. We took an eight amino acid sequence portion of the λ repressor protein and calculated the EML of its gene with and without the effects of AAVAR and CD. The EML for the occurrence of an invariant coding sequence for the 8-AA sequence (*58*) was approximately 2.8×10^{14} bases (Figure 1A). With CD, the EML was markedly reduced to approximately 2.7×10^{10} bases (Figure 1B). When AAVAR was introduced in addition to CD, the EML was drastically reduced to approximately 5,840 bases (Figure 1C).

The EML of a 400 AA protein with CD and AAVAR can be estimated based on an average CD of 3.2 codons per AA and average AAVAR of 16 AA per sequence position (see below): $EML_{400AA\ sequence} = 1/(CD * AAVAR/64)^{400} = 1/(3.2 * 16/64)^{400} = 10^{39}$ bases, which is immensely shorter compared to the EML of the invariant 400 AA sequence (10^{720} bases, see above).

Combined effect of CD, AAVAR, and gene splitting

Using the above principles, the EML of any exon is: $EML_{exon\ i} = 1/(p_{AA1} * p_{AA2} * \ldots * p_{AAni})$, where p_{AAi} is the probability of the i^{th} AA position.

In the simplest case when there is no CD and AAVAR, $EML_{exon\ i} = 1/p^{ni}$, where p is the probability of a given AA position or simply $EML_{exon\ i} = 64^{ni}$, and where the $exon_i$ contains n_i AAs.

When CD and AAVAR exist, the probabilities of the AA positions in the i^{th} exon are not equal and can be expressed

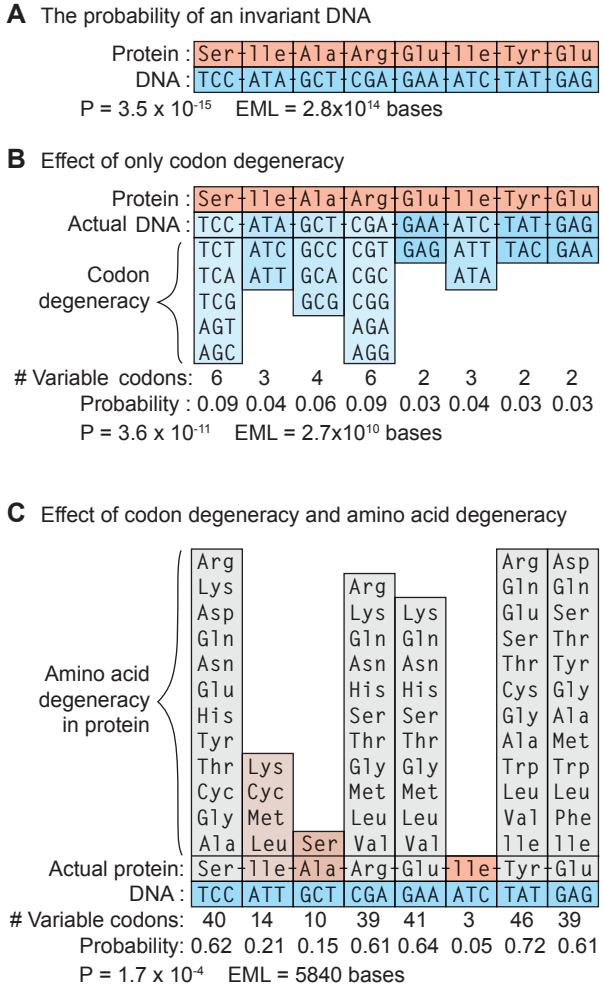


Figure 1. | CD and AAVAR increase the probability of a gene occurring in random DNA. (A) The probability and EML for an 8-AA sequence portion of the λ repressor protein with no CD or AAVAR, (B) with CD, and (C) with both CD and AAVAR. The degenerate codons for each of the variant AAs at each position are not shown in C. (P = Probability; EML = Expected Mean Length).

as $p_{ik} = (CD*AAVAR_{ik})/64$, where p_{ik} is the probability of the k^{th} position in the i^{th} exon and $(CD*AAVAR_{ik})$ represents the number of all possible codons on the k^{th} position in the i^{th} exon. In this case,
$EML_{exon\ i} = (64^{ni}) / (CD*AAVAR_{i1})*...*(CD*AAVAR_{i}n_{i})$.

It follows that the length of the longest exon, when significantly longer than the length of the next longest exon (which is true in most intron rich genes; see Supplementary Information Figure 2) becomes the primary determinant of the EML of the gene. Thus the EML of the longest exon will approximate the EML of the split gene.

$EML_{split\ gene} = EML_{exon\ 1} + ... + EML_{longest\ exon} + ... + EML_{exon\ n}$
$= EML_{longest\ exon} + \Sigma_{i \neq longest}\ EML_{exon\ i} = EML_{longest\ exon} + S$

$EML_{split\ gene} / EML_{longest\ exon} = 1 + S / EML_{longest\ exon} = 1 + \varepsilon$,
where ε is the approximation error and is close to zero (i.e., in the case of 40 exons with the longest exon at 100 bases and the second longest exon at 80 bases), $\varepsilon < 10^{(-12)}$.

Thus, $EML_{split\ gene} \approx EML_{longest\ exon}$

To illustrate the reduction in the EML of a DNA sequence coding for extant domains by the combined effect of CD, AAVAR and splitting the coding sequence, we used the variable sequence of a known domain available in the PFAM database (PFAM ID: PF00753, See Table 1). We split the AA sequence (100 AA) at multiple locations such that the length of the longest split sequence was successively reduced by 10 AAs or 30 bases. The expected EMLs for the un-split coding sequence and each of the split coding sequence configurations were computed using the above equations. Whereas the un-split gene (300 bases) has an EML of 1.1 x 10^{10} bases, a split gene with short-exons (30 bases) has an EML of just 770 bases (Table 1).

An empiric demonstration of the above theoretical calculation was also performed. Twenty-five random DNA sequences (each four billion bases) were computer generated. Each was searched for the presence of split coding sequences (the un-split 300-base sequence does not occur within this length) using the *Indigenous Gene Search* (IGS) algorithm (see Methods) that we developed. The search began at the first position of the sequence and terminated upon the first occurrence of the coding sequence. The length of DNA required to achieve the coding sequence was recorded and averaged over 25 iterations. The length of the split gene (experimental EMLs) and the predicted EMLs showed significant concordance (Table 1).

Number of splits (split-lengths in bases)[a]	Longest chain length (bases)	Predicted EML (bases)[b]	Experimental EML (bases)[c]
1 (300)	300	1.1 x 10^{10}	-
2 (30, 270)	270	4.4 x 10^{8}	1.7 x 10^{8}
3 (30, 30, 240)	240	9.1 x 10^{7}	8.0 x 10^{7}
4 (30, 30, 30, 210)	210	9.3 x 10^{6}	9.7 x 10^{6}
5 (30, 30, 30, 30, 180)	180	6.5 x 10^{5}	5.6 x 10^{5}
6 (30, 30, 30, 30, 30, 150)	150	150000	200000
7 (30, 30, 30, 30, 30, 30, 120)	120	53000	64000
8 (30, 30, 30, 30, 30, 30, 30, 90)	90	15000	15000
9 (30, 30, 30, 30, 30, 30, 30, 30, 60)	60	7300	6700
10 (30, 30, 30, 30, 30, 30, 30, 30, 30, 30)	30	770	770

Table 1. | Splitting a gene into exons drastically reduces the EML for the occurrence of a coding sequence of a protein (analysis of example protein PF00753 shown, AAVAR = 17 AA).
[a] Lengths of the split coding sequences for the protein sequence (exons)
[b] Sum of EMLs for each of the split coding piece for the specified split arrangement
[c] Average length of the random DNA in which the split gene occurred (over 25 iterations)

Split genes for complex multi-domain proteins

The above analysis was carried out for a single domain protein. We also wanted to know whether split genes encoded in random sequence could explain the origin of multi-domain proteins. To test this hypothesis, we designed five unique proteins (233–371 AA long) containing multiple extant domains from a set of 26 complex domains (PFAM database, Supplementary Information Table 1)**,** each with a different structure, function, sequence and length. For example, a DNA-binding domain (A), a kinase domain (F), a lactamase domain (K), a GAF domain (P), an OB-fold nucleic acid-binding domain (U), and a phosphotransferase domain (Z) were combined into a single protein (AFKPUZ). Each of these proteins was split arbitrarily without consideration to the number or positions of the splits, with the exception of the longest segment being ≤80 AA. The EML of the split gene for each of these hypothetical multi-domain proteins was calculated as described above (Table 1). Each of these split genes was searched in the same computer generated random DNA sequence of length 5 x 10^9 bases (which is ~100 x EML for the 80AA split; see Methods) over 100 iterations. The predicted and experimental EMLs were highly concordant and were immensely shorter than those of unsplit genes (Table 2). For instance, a contiguous gene for the 298 AA protein CHMRW would require 1.5 x 10^{30} bases of random sequence to occur once on average (with CD and AAVAR), whereas a split gene would occur in mere 1654 bases.

Genes for proteins with common domains and sequence repetitions

The previous experiment was repeated with the addition of extra domains, common to each of the proteins. For example, we introduced four common domains [integrase core domain (W), EF hand (X), phosphotransferase enzyme (Y), and leucine rich repeat (Z)] into each of the proteins, generating unique proteins that have certain domains in common. The predicted and experimental EMLs were consistent (Table 2). The locations of the genes with common domains in the random sequence were distinct (not shown) indicating that they were able to arise independently, without the need for a common ancestral domain.

LBE assumes that any sequence with internally repeated sequences (e.g., albumin, fibrinogen, collagen) evolved by sequence duplication from an original set of genes (*2-3, 60-62*). To determine whether split genes coding for proteins with sequence repetitions occur within random sequence, we produced repeated sequences from the protein CHMRW by repeating the unique domains (CCCCC, HHHHH, MMMM, RRRRR, and WWWWW) (Supplementary Information Table 1)**.** The EMLs of these genes were similar to those of genes without sequence repetition (Table 2). Experiments with repeated sequences from each of the model proteins shown in Table 2 produced similar results. Split genes coding for collagen with very short 3-AA repeats (Gly-X-Pro), believed to be the product of a billion of years of evolution, were also found to occur intrinsically in random sequences (Table 2).

Split genes for extant proteins

Though we used model proteins constructed from complex extant domains in above studies, we also validated our studies by using the domain sequences from PFAM in the same order as found in extant proteins. The results supported our findings that split genes for complete multi-domain proteins do indeed occur within random DNA sequence (Table 2). Though the average AAVARs of domains in these proteins were relatively low in the PFAM database (due to inadequate sampling), the EMLs of these genes were vastly lower than those of the un-split, contiguous genes.

Length of split genes found in random DNA

An EML represents the average length of a DNA sequence in which one copy of a gene can be found. While the experimental EML does represent one copy of a gene, shorter versions of the gene are likely to exist anchored to and including the longest exon. Once the longest exon is discovered in a random DNA sequence, many copies of the shorter exons will usually also be found at this location. Therefore, the expected length of the gene in a random sequence is essentially determined by the sum of the EMLs of the remaining exons, with the second longest exon being the primary determinant. In the phrase analogy, the EML of the phrase is the sum of the EMLs of all the words except the word NOT:

$EML_{phrase} = EML_{TO} + EML_{BE} + EML_{OR} + EML_{TO} + EML_{BE}$

$EML_{phrase} = 26^2 + 26^2 + 26^2 + 26^2 + 26^2 = 3380$ characters (compared to 20280 characters for the split phrase including the word NOT).

In addition, the length of the gene tends to reduce drastically with more iterations of the experiment. Based on probability theory, the formula for cumulative distribution function for exponential distribution is given by:

$$p_i = Pr\,(L_i < x_{pi}) = F_i(x_{pi}) = 1 - \exp\,(-\,x_{pi} / \lambda_i)$$

For the i^{th} exon, the $EML_{exon\ i} = \lambda_i$ and the equation can be rewritten in the form:

$$x_{pi} / \lambda_i = -\ln\,(1 - p_i)$$

For sequences no longer than $x_{pi} = 0.01*\lambda_i$, the probability $p_i = 1 - \exp\,(-0.01) = 0.01$. Therefore, 1 in 100 iterations will be shorter than $0.01*\lambda_i$. Similarly for $x_{pi} = 0.001*\lambda_i$, the probability $p_i = 1 - \exp\,(-0.001) = 0.001$. Therefore, 1 in 1000 iterations will be shorter than $0.001*\lambda_i$.

The above equation shows that 10% of EMLs will be shorter than approx. 0.1 times the mean waiting interval; and 1% of EMLs will be shorter than approx. 0.01 times the mean waiting interval; and so on. Therefore, after 10 iterations, the shortest phrase will be ~338 characters long; after 100 iterations, ~34 characters. The same is true for split genes. When the second longest exon was 230 bases (average in random genes and human genes; See Methods), the EML of the split gene was ~ 18 x 10^6 bases. After 1000 iterations, a ~ 7500 base version of the same gene was found.

Table 2. | Occurrence of split coding sequence for complex proteins containing multiple domains in random DNA. Model proteins were constructed using different combinations of 26 domains (shown in Supplementary Information Table 1) and split into short segments (≤80 AA). A computer generated random DNA (four billion bases in length) was searched for the occurrence of the split genes coding for these model proteins in two forms: the long split gene (forward search) and the short split gene (reverse search) over 100 iterations (See Figure 2).

Model protein sequence	AA (base) sequence lengths	Mean AAVAR	Number of exons (Split lengths in AAs)	Predicted EML for contiguous sequence (bases)	Predicted EML for split sequence (bases)	Experimental EML for split sequence (bases)	Shortest split gene length (bases)
Protein sequences containing unique domains							
AFKPUZ	238 (714)	15.8	**14** (8|7|7|8|10|8|8|8|15|40|10|80|15|14)	2.27×10^{38}	1.17×10^{7}	1.3×10^{7}	43573
BGLQV	233 (699)	17.2	**10** (10|15|10|15|40|15|80|20|10|18)	1.22×10^{22}	5.22×10^{6}	5.1×10^{6}	971
CHMRW	298 (894)	17.1	**13** (10|10|20|25|20|40|10|80|13|30|15|10|15)	1.54×10^{30}	7.2×10^{6}	7.1×10^{6}	1654
DINSX	292 (876)	17	**14** (15|25|10|25|25|12|10|10|80|10|40|10|10|10)	3.88×10^{28}	2.0×10^{6}	1.8×10^{6}	1293
EJOTY	371 (1113)	17.1	**18** (10|10|25|30|10|10|10|10|40|10|10|80|20|15|20|25|25|11)	1.45×10^{36}	3.9×10^{6}	4.7×10^{6}	20040
Protein sequences containing unique and common domains (W, X, Y, Z)							
AFWKPUXYZ	370 (1110)	16	**22 (**10|10|10|10|10|10|10|10|15|30|10|15|60|15|40|10|10|15| 10|30|20|10**)**	1.72×10^{54}	5.7×10^{6}	4.7×10^{6}	43466
BZGLYQWVX	374 (1122)	16.9	**20** (10|15|10|20|15|20|14|10|10|20|10|20|20|40|10|60|30|20| 10|10)	1.67×10^{39}	1.1×10^{7}	1.1×10^{7}	1929
CYHZMRWX	376 (1128)	16.9	**17** (15|22|15|22|10|10|10|20|20|80|10|40|25|10|20|30|17)	1.12×10^{40}	1.7×10^{7}	2.0×10^{7}	7697
ZDIWNSXY	440 (1320)	16.9	**20** (10|10|15|25|10|25|25|30|20|15|20|40|15|80|20|15|20|20| 10|15)	1.67×10^{45}	1.5×10^{5}	1.3×10^{5}	5749
XEJZOWTY	455 (1365)	17	**26** (10|20|10|25|10|20|15|10|15|15|10|30|10|20|15|15|30|10| 60|10|25|20|15|10|15|10)	8.41×10^{45}	5.6×10^{5}	5.8×10^{5}	3221
Protein sequences containing repeated domains							
CHMRW	298 (894)	17.1	**13** (11|20|15|20|12|40|10|80|20|10|25|15|20)	1.54×10^{30}	1.0×10^{6}	9.9×10^{5}	3641
CCCCC	65 (195)	18.8	**3** (15|20|30)	4.21×10^{5}	2.4×10^{2}	2.7×10^{2}	195
HHHHH	335 (1005)	17.4	**17** (10|15|10|10|10|10|80|15|40|10|15|25|20|10|30|15|10)	1.47×10^{29}	2.4×10^{6}	1.9×10^{6}	1197
MMMMM	615 (1845)	17.1	**32** (10|15|20|10|15|30|15|25|10|25|15|20|20|40|10|15|60|10| 15|30|15|30|15|10|10|15|10|30|10|15|30|15)	1.60×10^{57}	1.7×10^{5}	1.7×10^{5}	11480
RRRRR	160 (480)	16.8	**7** (15|10|40|10|60|10|15)	3.86×10^{19}	8.9×10^{5}	8.3×10^{5}	650
WWWWW	315 (945)	16.6	**15** (10|15|10|25|10|30|10|60|10|40|15|15|15|20|30)	6.39×10^{38}	6.2×10^{6}	6.2×10^{6}	9154
Collagen	60 (180)	7.5	**8** (7|8|7|8|7|8|8|7)	2.60×10^{34}	1.8×10^{5}	1.8×10^{5}	32160
Multidomain Uniprot Proteins							
Serine/threonine-protein phosphatase with EF-hands 1 (PPEF1 / O14829-1)	206 (618)	11.3	**37**(4|5|8|4|5|9|8|7|5|7|10|6|4|6|5|8|5|6|7|3|3|3|4|7|8|4|5|3|7|3|5|8|4|3|7|6|4)	6.09E+76	2.14E+06	2.14E+06	62366
LRP2-binding protein (Lrp2bp / Q569C2)	88 (264)	16.5	**8**(8|14|10|14|20|8|7|7)	6.49E+09	4.68E+02	4.89E+02	275

The above phenomenon was illustrated in a simple example. We searched for a split gene encoding a 17 AA portion of the λ repressor protein, and marked the locations of each of the exons in random sequence until the complete gene was found (Figure 2). We used the reverse search algorithm that we developed to conduct this analysis (Figure 2A; see Methods). Whereas the length of the split gene found in the forward search was long (1886 bases), the reverse search yielded a short gene length (83 bases, Figure 2B). Though the exon D was the longest (7 AAs), the exon B (2 AAs) was the least probable due to its very low AAVAR, and therefore its EML was the largest.

The same pattern was observed with a variation of this experiment, in which three exons were examined. The forward search resulted in a gene of 70,766 bases, however the reverse search still resulted in just 495 bases (Figure 2C).

The frequency distribution of the lengths of the forward and short genes over 1000 iterations displayed that shorter genes occurred far more frequently than longer genes in both cases (Figures 2D and 2E corresponding to Figures 2B and 2C). We applied this principle to all the model proteins for finding the short split gene at the location of the longest (least probable) exon (Table 2). The length of the shortest split genes after 100 iterations was much shorter than the EMLs of the split genes, and was immensely shorter when compared to the EMLs of their corresponding contiguous genes. For instance, the contiguous coding sequence for the protein BZGLYQWVX (374 amino acids; 1122 bases) requires 1.67 x 10^{39} bases for its chance occurrence (with CD and AAVAR), whereas the short split gene found surrounding the longest exon was 1929 bases.

Genes for numerous proteins occur within the same finite random DNA

The exons of most intron-rich extant genes, and genes predicted in random DNA, are limited to a finite length as predicted by ROSG (54-57). Therefore, irrespective of the number of exons (e.g., 5 or 100), or the length of the protein sequence it encodes (e.g., 500 or 10,000 AAs), the split gene should occur within essentially the same finite random sequence provided that the sequence is of sufficient length. We have previously shown that ~600 bases is the statistical maximum for any exon (53-57, A. Bhasi, et al, accompanying paper). Therefore, a random sequence of DNA with length, for example, of 10 x $EML_{600\text{-base-exon}}$ should be expected to contain all extant intron-rich genes in that sequence with greater than 99% certainty.

To test this prediction, we conducted 41 gene-search experiments (Table 2) within the same computer-generated random sequence of four billion bases, which is 100 times the EML for the longest exon (80 AA, 240 bases). The shortest split gene for each of the proteins after 100 iterations (Table 2) was found within the first 14 million bases of the random sequence (Figure 3). Thus, split genes (with exon lengths ≤240 bases) for virtually any given protein sequence should occur within this same finite random sequence -- akin to the fact that a random stream of English alphabets of length 10^{16} characters should contain any sentence with the longest word of 10 characters (100 x 26^{10} characters) from any book ever written.

As mentioned above, ~600 bases is the statistical maximum length for any exon. In addition, the second longest exon is much

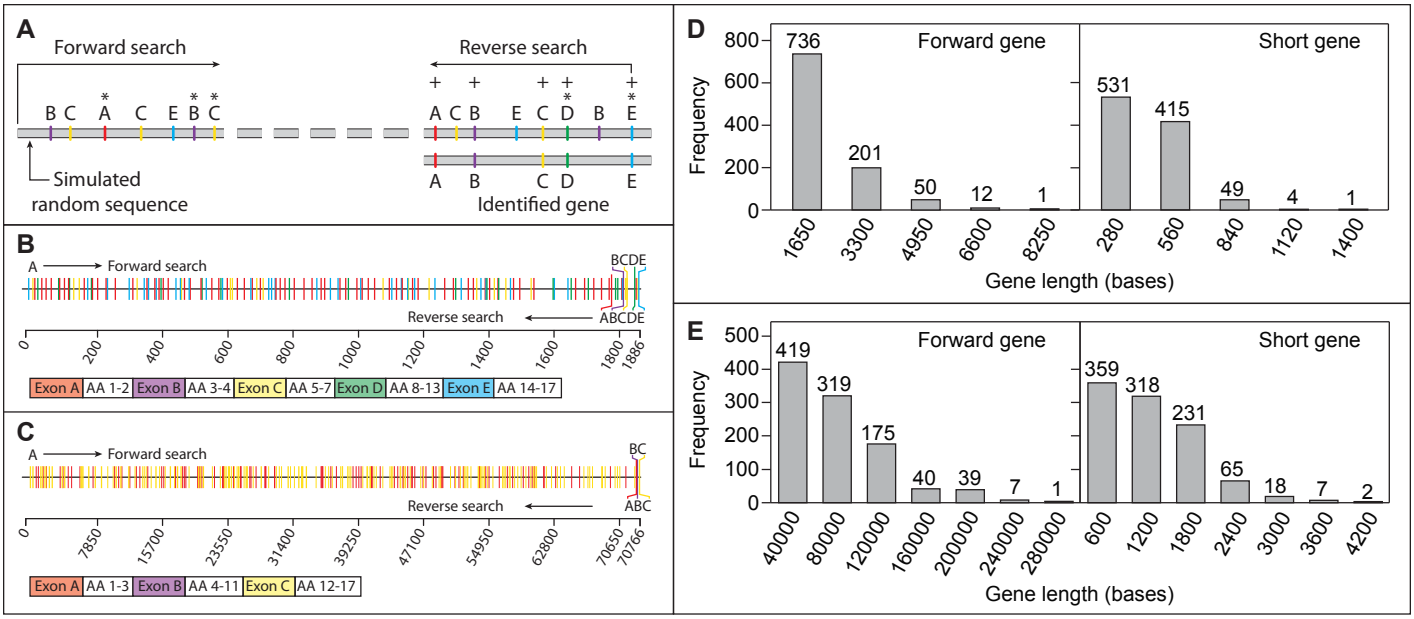


Figure 2. | The increase in gene probability after splitting. (A) The reverse gene search algorithm (See Methods). Exons A-E found in the forward search are marked with an *, and those in the reverse search with a +. (B) All the occurrences of each of the five exons of the 17-AA segment of the λ repressor found in random DNA are shown in respective colored vertical lines on the horizontal line representing the random DNA sequence. The first occurring five exons found in the forward search are shown above the lines and those in the reverse search are shown below the long lines. The length of the "forward" ABCDE pattern was 1886 bases, and the "short" ABCDE was 83 bases. (C) A forward and reverse search for the three exon splits of the same 17-AA λ repressor segment. The long gene was 70,766 bases and the short gene was 495 bases. (D&E) The distribution of the frequency of the lengths of forward and short split genes over 1000 iterations.

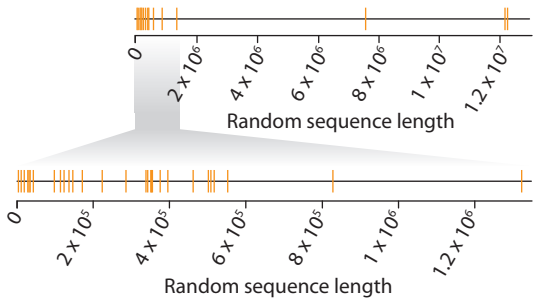


Figure 3. | Unique genes for numerous complex proteins occur within the same finite random DNA sequence (4 billion bases). The occurrences of the shortest split genes, among 100 iterations of each of the 41 protein sequences (Table 2 and their repetitive sequences), were marked to scale as vertical line segments on a horizontal line representing the random DNA sequence. All of the shortest genes occurred within the first 14 million bases. The portion of DNA for the first 39 genes is magnified. The gene names are not shown for convenience.

shorter (~230-300 bases) in extant intron-rich genes (*53-57*, A. Bhasi, et al, accompanying paper). An extension of the above analyses shows that a random DNA sequence of ~10^{12} bases should contain split genes with a maximum exon length of ~600 bases, encoding any given protein with an average AAVAR of 16 AA/position (Supplementary Information Figure 3). It would require approximately 10,000 - 100,000 iterations (a random DNA length of ~10^{16} bases, or ~1 µg of DNA) to obtain a smaller version of this gene (10^3–10^6 bases; calculated based on the average length of the second longest exons), which is found in modern intron-rich genes. This one microgram of DNA is sufficient to contain the split genes encoding every protein in the biota.

DISCUSSION

Origin of biological information in split pieces

The pre-biotic chance origin of proteins has never been explained, as the analyses have been based on prokaryotic contiguous genes that are simply improbable of occurring in any amount of random genetic sequences (*1-7*). Therefore, the origin of biological information and the origin of life is still a great question, often answered by theories concerning an improbable chemical accident in the prebiotic pool (*4*), or foundations in outer space (*63*). Even after the knowledge that eukaryotic split genes may have been the very first genes became widespread (*39, 53-57, 64-65*), there has been no attempt to analyze the origin of eukaryotic genes and proteins from random genetic sequences other than the ROSG model.

Our study shows that eukaryotic split genes could have originated by chance in pre-biotic chemistry, which then led to prokaryotic genes. Splitting the coding sequence of a gene into several short pieces and separating them with introns can achieve almost any level of biological complexity from random sequences. That random sequences existed in the prebiotic pool is an assumption of ROSG as well as LBE. The proteins of the biota are thought to be the products of a billion years of evolution. In contrast, our analyses show that the coding sequence for a complex protein, including its structural and functional regions, occurred within random genetic sequences and did not undergo major evolution. In addition, an important indicator of the random origin of proteins is the probability that a complex protein sequence occurring in split gene form is the same as that of a random AA sequence.

LBE assumes that:

1) complex multi-domain proteins evolved from simple ancestral domains,
2) common domains among proteins were passed on from a single domain in the first life form, and
3) sequence repetitions were duplicated from precursor sequences (*2-7*).

However, based on our findings, these are incorrect and not necessary to explain observed facts. The finding that multi-domain proteins with common domains occurred independently in random sequence explains why the same domains are present in proteins without orthology (*8-11*). Using the Exon-Domain comparative analytical tools (ExDom), we found that the exon-intron structures of the region of genes coding for common domains in entirely different proteins were completely unique, indicating their independent origins (PS Unpublished). Several examples of such independent proteins with highly similar 3D structure lacking of sequence similarity or orthology exist (*8-11, 66*). Furthermore, under the ROSG model, sequence repetition does not require gene-duplication events, because a high level of redundancy is possible among independent genes due to structural constraints imposed by functional constraints under similar biochemical situations (*53*). We found that the Gly-X-Pro sequence repetition in collagen, for example, was produced in random sequence. The results are consistent with the recent findings that gene-duplication, the only mechanism purported to have evolved new gene functions, cannot evolve entirely new genes and new functions and can at best be responsible for sub-functionalization of pre-existing functions (*67-71*). Recent findings also indicate that the evolution of different extant protein domains from an assumed few primordial protein structures, and the way in which new structural frameworks evolve via simple mutations, cannot be explained (*72-75*). We also found that the lengths of the split genes coding for multi-domain proteins, and proteins with common domains or repetitive sequences that occur in random sequence are consistent with split gene lengths in extant intron-rich genomes. In addition, ROSG requires that protein size be proportional to the length of the split gene and the number of exons and introns in the gene, and this idea was confirmed in this study (*54*).

Pre-biotic origin of complex eukaryotic genomes

Our study indicates that a complex eukaryotic genome could have self-assembled from the vast pool of pre-biotic split genes by the same pre-biotic self-assembly mechanisms that are currently believed to have brought forth the genome of the first primitive life on Earth (*4, 6, 76*). Under ROSG, regulatory sequences such as promoters with multiple binding sites required for complex genetic networks, also occurred along with protein-coding genes (A. Bhasi, et al, accompanying paper).

Multiple eukaryotic genomes could have also arisen at the same time due to the sheer abundance of split genes and their regulatory sequences. They may have evolved pre-biotically as efficiently as unicellular genomes if there was no substantial difference in the complexity of the genes, proteins or regulatory sequences between these life forms. Recent discoveries of the equal complexity of metazoan genomes from the basal trichoplax and sea anemone to the 'highest' eukaryotes such as the human (*77*) support this idea. Further evidence for the multicellular genome origins from a pre-biotic pool of eukaryotic genes resides in recent findings that entirely new genes are present in life forms without any precursors in their expected ancestors, and that the genes of the eukaryotes are distributed in a mosaic manner across genomes (*78-83*), (P. Senapathy, et al, accompanying paper).

The universal occurrence of the 4-character DNA alphabet, 20-character protein alphabet, genetic code, codon degeneracy and AA variability in most or all extant life forms is thought to be due to their frozen accident within the very first life form and their propagation in all life forms through evolution. However, according to ROSG, these molecular entities could have been established within pre-biotic chemistry as the stochastically best possible combination for encoding maximum biological information within a minimum amount of random DNA, and were thus used in the assembly of all genomes.

Origin of eukaryotic introns

The present findings support the ROSG model for the origin of introns and the split structure of eukaryotic genes. We have previously shown that intrinsic open reading frame length constraints severely restricted the exon length and forced the coding sequence to be split (*53-57*). Furthermore, the low probability of the combination of the splicing signals and the split biological informational pieces must have caused the introns to be exceptionally long. In addition, the location of the stop codons exactly at the ends of exons as parts of splice signal sequences also supports ROSG (*53-57* and A. Bhasi, et al, accompanying paper).

We have shown in a separate study that complete split genes, containing regulatory elements, splicing signals, exons and introns, must have occurred in pre-biotic DNA at ample frequency (A. Bhasi, et al, accompanying paper). Different combinations of splice signals and coding sequences that indigenously occurred within random sequence may have enabled a large repertoire of alternatively spliced gene variants to descendent genomes (*84-85*). Along with our other findings (*53-57*, A. Bhasi, et al, accompanying paper), the present analysis demonstrates that protein coding and ncRNA genes as well as the other regulatory elements required for the pre-biotic evolution of a complete genome could occur within a finite length of random sequence.

There appears to be a general assumption in biology that the origin and diversity of life followed a simple-to-complex pathway. Our work, however, shows that it was far more probable for the structurally complex eukaryotic genomes to have originated first and then "reduced" into simpler genomes in the pre-biotic system (*53-57*, P. Senapathy, et al, and A. Bhasi, et al, accompanying papers). Thus, intron-poor genes in small genomes (e.g., *C. elegans, P. falciparum, C. merolae, or yeasts*) may be the result of intron loss in the pre-biotic system. Arabidopsis, trichoplax and sea anemone, which contain intron-rich genes with short introns, may be the result of reduction in intron length (*54-57*). Based on ROSG, such intron, gene or genome reductions could have happened within the pre-biotic system rather than in evolving organisms. Prokaryotes may be the end result of complete loss of introns from full-fledged intron-rich genes.

In addition, there are no primitive proteins in the biota; all proteins are more or less equally complex. The scenario of the protein world does not show a gradation of simple to complex proteins. Furthermore, there is no theoretical framework under the LBE model for the pre-biotic origin and evolution of even a primitive protein (*1-7*).

Origin of eukaryotic nucleus and sub-cellular structures

As the LBE model states that the first life form was a bacterium like organism, biologists correspondingly proposed that the eukaryotic cell had evolved from simpler cells such as bacteria, archaeabacteria or their combinations by means such as endosymbiosis or phagocytosis (*25-26, 48-52, 86-91*). It has however emerged from post-genomic analysis that a eukaryotic cell or genome could not have evolved from a prokaryotic cell or genome, but that the *vice versa* may be true. Even after decades of research, no consensus framework for the evolution of a eukaryotic cell from bacterium-like cells has emerged (*25-26, 48-52, 86-91*).

Supported by the post-genomic data, our study suggests that the first life form was a complex eukaryotic cell with a full complement of sub-cellular structures. By demonstrating the inherent occurrence of complex split genes in random DNA, ROSG shows that the set of proteins for any complex eukaryotic cellular structure could have originated directly in the pre-biotic molecular system. A strong selective pressure would have also existed for a nuclear-cytoplasmic division in the first cell to avoid the molecular confusion between primary RNA transcription, splicing and mRNA translation, as well as to prevent competition between the spliceosome and ribosome for the binding of RNA molecules (*53-57*). In addition, the regulatory sequences for the network of genes required for the construction of the complete eukaryotic cell would have occurred in the pre-biotic genetic sequences (P. Senapathy, et al, accompanying paper).

Our findings thus show that any split gene that can encode complex proteins which form the structure of the spliceosome or any other eukaryotic cellular organelle, act as regulatory DNA binding proteins, function as a master control protein in the development of organs and appendages, or serve any other purpose as needed for the functioning of an organism, can occur within one microgram (~10^{16} bases) of DNA. The addition of structures such as splice signals would increase this amount to one milligram (~10^{19} bases) or so of pre-biotic random DNA.

Conclusion

This work does not claim to provide historical details of early evolution. We do not address the chemistry of prebiotic DNA materialization or the mechanism of the first spliceosome emergence. Rather, we have shown that biological information could have existed in split genes in random sequence, and that these genes could have been used in the self-assembly process to create countless eukaryotic genomes.

The ROSG model provides a stochastic, non-teleological mechanism for randomly arriving at a high level of biological complexity. By providing a solution to the long-standing problem of the origin of biological information, this study solves the origin of life and the origin of biological complexity, which faced fundamental problems when examined by the LBE model. By demonstrating the origin of intron-rich genes in pre-biotic genetic sequences, we provide consistent explanations for the origin of exons, introns, complex domains, multi-domain proteins, proteins with common domains and proteins with sequence repetitions. In effect, this study demonstrates that all of the complexity of life on earth can be traced back to the pre-biotic origin of biological information in split genes.

Methods

Datasets

Variable amino acid (AA) sequences for protein domains from the PFAM database, which is a large collection of protein families, were used; each was represented by multiple sequence alignments and hidden Markov models (HMMs). Release 22.0 of PFAM consisted of 9318 families, including 2,990,695 sequences. The annotation and seed alignment of all PFAM-A families (Pfam-A.seed) were downloaded from the PFAM FTP site (ftp://ftp.sanger.ac.uk/pub/databases/Pfam/releases/). The seed alignment file was parsed to generate a dataset of the 9318 variable sequence matrices of protein domains, each with a unique Accession number. Gap characters other than the single letters representing the 20 amino acids were omitted from the sequence matrix. The variable amino acid sequence for the λ repressor protein was used from *(58)*. A random sequence of four billion bases was generated and used throughout this work.

Codon degeneracy (CD) and amino acid variability (AAVAR)

Average CD

We searched for portions of the DNA sequence that coded for any one of the numerous variable amino acid sequences representing the protein/domain (e.g., λ repressor protein; Supplementary Information Figure 2) in the computer generated random DNA sequence. Accordingly, the degenerate codons for each of the variable amino acids present at a given amino acid sequence position were considered (e.g., Proline: CCT, CCC, CCA, CCG). The average CD for an AA was computed, by dividing the total AA coding codons *(61)* by the total number of amino acids in proteins *(20)*, to be 3.05 codons per AA.

Average AAVAR per sequence position

The variant residues at each position of the multiple aligned sequences of the homologous domains in the PFAM 'seed file' were grouped and referred to as the variable AA group. Thus, for the variable AA sequence of a given domain, there would be as many variable AA groups as the length of the domain sequence. The number of variable amino acids at each AA sequence position, and the total number of variable amino acids present within the complete protein domain were computed. The average AAVAR per AA sequence position was obtained by dividing the sum of all the AAVARs within the whole protein domain sequence by the length of the domain sequence.

The average AAVAR was found to vary widely among proteins. There were very few sequences within the PFAM seed alignment file for some homologous domains. As a high level of AAVAR is a basic property of proteins in general *(58)*, the higher end of the variability found in PFAM domains should represent the natural variability. Thus, we selected the 26 domain sequences with the highest AAVARs ranging from 16.0 to 19.5 AAs per sequence position (Supplementary Information Table 1). A very high AAVAR of 19.5 AAs per position may not occur in nature due to the possibility of the dependency of particular amino acids among different positions of a domain. However, it has been established in the literature that the AAVAR in a protein is very high, up to 90-95% *(58)*. Because this study considers the probability of the split gene coding for a given protein, rather than the contiguously coding gene, the probability of a given gene is still very high even with lower AAVARs of 15-16 AAs per position (75-80%) or lower.

CD in a random DNA sequence

The frequency of occurrence of each of the 20 AAs in biological proteins is found to be proportional to the number of degenerate codons (i.e., Met - 1, Arg - 6, Ala - 4) *(2, 53)* (Supplementary Information Figure 4). At each codon position with a degenerate codon in a random DNA sequence, a corresponding number of degenerate codons can code for the same amino acid. Thus the total number of variable codons for every 61 AA positions is 235, increasing the average codon degeneracy from 3.05 (61/20) to 3.85 (235/61). This phenomenon considerably increases the probability, and thus decreases the EML, of the gene coding for a given variable protein sequence in a random DNA sequence. However, in our calculations, this higher value of 3.85 would increase the average number of codons per AA position to more than 61 codons for the 20 AAs (20 x 3.85). The true average CD that can be used in these calculations is between 3.05 and 3.85 (e.g., 3.45) (Supplementary Information Figure 4). However, we have used a conservative value of 3.2 in our study.

The probability and EML of a variable protein within a random DNA

We developed an algorithm for computing the probability and EML of the coding sequence for a protein of given length, based on the average CD and AAVAR. For a given AA sequence, (as in Figure 1), the variable AAs at each position were grouped into a "variable AA group". Next, the sum of the number of degenerate codons for all of the different AAs within each variable AA group was computed. The probability of the variable AA group at each sequence position was computed by p_i = sum (degenerate codons for each residue within the variable amino acid group)/ 64. The probability of a sequence that codes for a protein with AA variations is computed using the formula $p_{protein} = p_1 * p_2 * \ldots p_n$, where p_i is the probability of the variable AA group at each AA position, and n is the number of variable AA groups in the protein (i.e., the length of the protein). The EML of the random DNA for the chance occurrence of the protein sequence of a given length n was computed using the formula $(1/p_{protein})$.

Using the above procedure, the probability and EML for each of the 9318 protein sequences were computed. Then, for each domain length represented in PFAM, the protein domain with the highest average AAVAR was listed (Supplementary Information Table 2). From the list of 666 protein domains, we chose 26 domains with the highest average AAVAR (listed in Supplementary Information Table 1). The average AAVAR of these proteins, V (17.1 amino acids per sequence position), was combined with the average CD (3.05 codons per AA) to compute the average probability for an AA at any given position in a protein sequence (P = V x CD/64). The approximate probability and EML of any protein-coding sequence of a given length can be computed based on this value using the formulas

$$p_{protein} = (p_i)^n$$
$$EML = (1/p_{protein}),$$

where p_i is the average probability at a given sequence position, and n is the length of the protein sequence.

Searching for the split coding-sequence

We developed a computer program IGS (Indigenous Gene Search) designed to search for the consecutive occurrences of the split coding-pieces of a variable protein sequence (considering AAVAR & CD for each variable AA) skipping the random DNA sequences between the coding sequence segments. The steps in IGS algorithm are: 1) For every degenerate codon of every variable AA at a given sequence position of the protein domain, a match with the first codon of the random DNA sequence is sought. 2) If a match occurs, the search is continued with the second codon in the random DNA and the degenerate codons of the amino acids within the second variable AA group. If a match does not occur, step 1 is repeated with the next codon in the random DNA sequence. 3) Steps 1-2 are repeated until the end of the first protein split. 4) When there is a complete match of the random sequence to the codons of the first split segment of the AA sequence, the search for the second split segment is repeated with Steps 1-3, starting with the next codon within the random DNA. 5) The random sequence is ignored until the complete match for the second split segment has been obtained. 6) The search is continued with the third and subsequent split segments until the end of the protein is reached.

The split segments of the coding sequence representing each split in the protein sequence were akin to exons and the intervening random sequences were akin to introns. In this study, we used the term exon to denote the split coding sequences corresponding to the split protein sequences. The program computed the length between the occurrence of the first exon and the last exon in the random DNA (forward "split-gene"). A gene thus found would code for any one of the variable sequences of the protein that were used as the input variable sequence. The average length of the random sequence from the start of the search to the end of the random sequence in which the complete split gene was found (over 100 iterations) should represent the true EML of the split gene. The true EML was also predicted by summing the EMLs of each individual protein sequence split, computed based on actual AAVARs at each sequence position and the CD (see above). The experimental EML obtained in gene-search experiments (average of multiple iterations) should match the predicted EML, which serves as a control.

Reverse searching for the short split gene

The Reverse Split Gene Search algorithm was developed for identifying the "shortest" split-gene for a given coding sequence in a random DNA (Figure 2A). The forward search (IGS, see above) first sequentially searches for and identifies exons (A through E that code for the protein splits 'A' through 'E') in the random DNA sequence, ignoring the intervening random DNA sequences between the occurrences of the consecutive splits. After the last exon has been located, the algorithm searches the random sequence in the reverse direction and located the shortest occurring ABCDE exon pattern. While the forward search ensures at least one occurrence of each of the splits of the complete split gene (forward split gene), the reverse search locates the shortest gene with the shortest intervening sequences. The finally located shortest ABCDE pattern represented the shortest split-gene for the given coding sequence (short split gene).

The average length of the short split genes obtained in a statistically significant number of iterations (approx. 100 iterations) represents the EML for finding all of the exons of the gene surrounding the longest exon, once the longest exon has been located. We found this value to converge to the EML of the second longest exon, which is predictable using its AAVAR (see above).

Multiple iterations of the search for short split genes

The Reverse Split Gene Search program was iterated over a given number of times (e.g., 100) consecutively in a long random DNA for obtaining a set of 100 hits in the forward direction, and 100 hits in the reverse direction, for a given protein with a predefined number and length of splits (the total split lengths is equivalent to the length of the protein). The average lengths of the forward and short split genes were computed from the results of 100 iterations.

Plotting the frequencies of gene-lengths

The length of the random sequence from the start of the search up to the first occurrence of the ABCDE gene (the end of exon E) in the forward search, and the length of the short split gene (the length of the first shortest pattern of ABCDE in the reverse search), were computed for a given number of search iterations for the split-genes with different numbers of exon splits and different exon lengths for a segment of the λ repressor protein (Supplementary Information Figure 1). The negative exponential distribution (NED) of the frequency of split gene lengths indicated that the shortest gene lengths were the most frequent and the longer genes became rapidly less frequent (Figures 2D & 2E). This NED nature is applicable for any given sequence split into shorter sub-sequences in a random sequence *(92)*.

Creating model proteins with extant domains

The variable sequences of different extant domains having higher AAVARs in PFAM (Supplementary Information Table 1) were appended randomly in various combinations and were categorized under a) protein sequences containing unique domains, b) protein sequences containing unique and common domains and c) protein sequences containing repeated domains, respectively (Table 2). Each of the variable protein sequences thus created was split arbitrarily, keeping the length of the longest split below 80 amino acids (240 bases). The IGS algorithm located the occurrence of each split sequence within the random sequence, and computed the lengths of the forward and short split genes. We calculated the predicted EML (the sum of the EMLs of each split protein sequence) and the experimental EML (the average length of 100 forward genes by IGS) of the split gene coding for each protein, and compared them with the EML of the contiguous coding (i.e., without splitting) form of the gene sequence (Table 2).

Splice signals and regulatory sequences

As our primary focus was on the analysis of proteins, our search for split genes in random sequences did not include splice signals and regulatory elements. We have addressed the splice signals and regulatory sequences in a separate study (5-8 and A. Bhasi et al, accompanying paper). However, based on the high probability of the regulatory sequences and splice-signals (all of which occur with considerable sequence variations), there would be no qualitative difference in the results of the current study—with only a slight difference in the probability and EML of the split gene (protein) sequences with and without the inclusion of the splice signals and regulatory sequences. Thus, the probability of the model proteins may be only slightly reduced (and the EML slightly increased) by requiring the split-gene to contain all of the authentic structural features including promoters, splice signals, and polyA site sequences.

In constructing the multi-domain proteins, we used arbitrary domains from the PFAM database, giving primary importance to the AAVAR of the domain sequence. A number of domains were concatenated, such that the total length of the constructed protein was fairly long. Though these proteins did not represent the particular content or order of the multiple domains in actual biological proteins, they contain modern biological domains in complex proteins with sophisticated structure and function with AAVARs as provided in the PFAM database. These proteins have the standard characteristics of complex proteins within any proteome, such as the length, sequence complexity, and AAVAR. Furthermore, the actual variability of biological proteins may be slightly lower than those that are represented by the highest variable PFAM sequences due to a possible inter-dependency between the amino acids at different positions within a protein sequence *(58)*. However, we expect that the qualitative results of our study will be essentially the same even if we use domains with reasonably lower AAVARs. We also chose a few extant proteins all of whose domains were represented in the PFAM database. We created the models for these actual extant proteins by concatenating the different domains for these extant proteins in the same order as they occurred in the actual proteins.

Acknowledgements

We wish to thank Vinu Manikandan, Kanika Arora, Ashwini Bhasi, Phillip Simon, Jeffrey Mattox, Kavin Senapathy, and H. Adam Steinberg for their assistance in improving the manuscript, and Tomasz Wojciechowski for assistance with statistics and probability. We also wish to thank ArtforScience.com for help in preparing the figures.

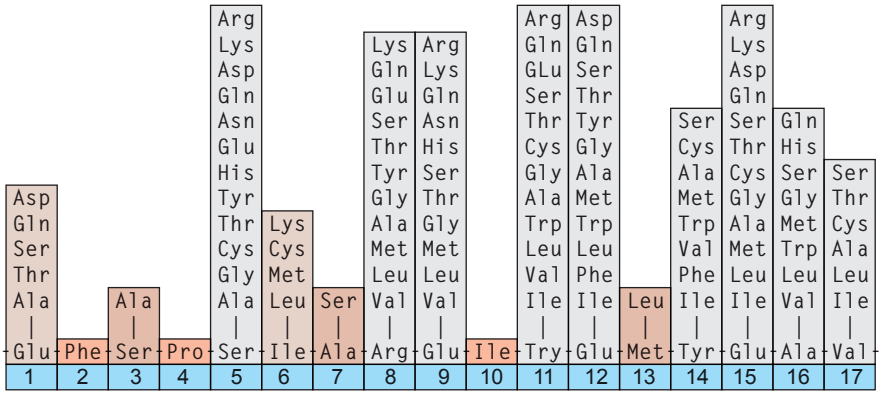
Correspondence should be addressed to P. S. (ps@genome.com)

References

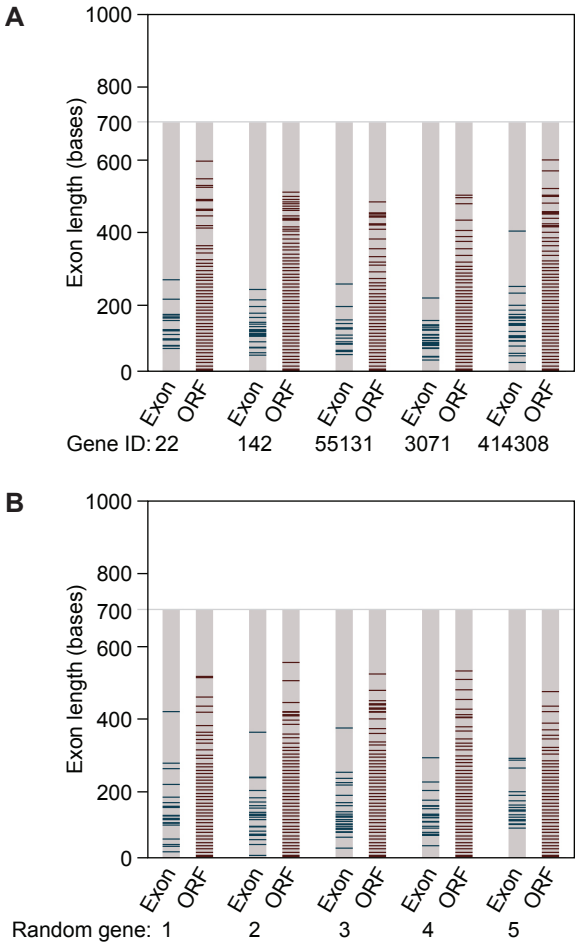
1. N. Tokuriki, D. S. Tawfik, *Science* **324**, 203 (Apr 10, 2009).
2. R. F. Doolittle, *Science* **214**, 149 (Oct 9, 1981).
3. R. F. Doolittle, *Trends Biochem Sci* **14**, 244 (Jul, 1989).
4. Kuppers. (The MIT Press, Massachusetts, 1989), pp. 59-61.
5. F. B. Salisbury, *Nature* **224**, 342 (1969).
6. J. Monod. (Vintage Books, New York, 1972).
7. H. P. Yockey. (Cambridge University Press, New York, 2005).
8. C. P. Ponting, R. R. Russell, *Annu Rev Biophys Biomol Struct* **31**, 45 (2002).
9. C. A. Orengo, J. M. Thornton, *Annu Rev Biochem* **74**, 867 (2005).
10. R. L. Marsden *et al.*, *Philos Trans R Soc Lond B Biol Sci* **361**, 425 (Mar 29, 2006).
11. G. Caetano-Anolles, M. Wang, D. Caetano-Anolles, J. E. Mittenthal, *Biochem J* **417**, 621 (Feb 1, 2009).
12. C. G. Kurland, B. Canback, O. G. Berg, *Biochimie* **89**, 1454 (Dec, 2007).
13. N. Glansdorff, *Mol Microbiol* **38**, 177 (Oct, 2000).
14. M. Wang, L. S. Yafremava, D. Caetano-Anolles, J. E. Mittenthal, G. Caetano-Anolles, *Genome Res* **17**, 1572 (Nov, 2007).
15. S. Yang, R. F. Doolittle, P. E. Bourne, *Proc Natl Acad Sci U S A* **102**, 373 (Jan 11, 2005).

16. B. Labedan *et al.*, *J Mol Evol* **49**, 461 (Oct, 1999).
17. G. Caetano-Anolles, D. Caetano-Anolles, *Genome Res* **13**, 1563 (Jul, 2003).
18. P. Forterre *et al.*, *Biosystems* **28**, 15 (1992).
19. R. L. Sherrer, P. O'Donoghue, D. Soll, *Nucleic Acids Res* **36**, 1247 (Mar, 2008).
20. P. Alifano *et al.*, *Microbiol Rev* **60**, 44 (Mar, 1996).
21. A. Habenicht, U. Hellman, R. Cerff, *J Mol Biol* **237**, 165 (Mar 18, 1994).
22. N. Benachenhou-Lahfa, P. Forterre, B. Labedan, *J Mol Evol* **36**, 335 (Apr, 1993).
23. B. Labedan, Y. Xu, D. G. Naumoff, N. Glansdorff, *Mol Biol Evol* **21**, 364 (Feb, 2004).
24. O. Kandler, in *Early Life on Earth*. (Columbia University Press, New York, 1994), vol. 8, pp. 152-160.
25. A. Poole, D. Jeffares, D. Penny, *Bioessays* **21**, 880 (Oct, 1999).
26. A. M. Poole, D. C. Jeffares, D. Penny, *J Mol Evol* **46**, 1 (Jan, 1998).
27. D. Penny, A. Poole, *Curr Opin Genet Dev* **9**, 672 (Dec, 1999).
28. N. Glansdorff, Y. Xu, B. Labedan, *Biol Direct* **3**, 29 (2008).
29. L. A. Katz, *Int J Syst Evol Microbiol* **52**, 1893 (Sep, 2002).
30. P. Lopez, P. Forterre, H. Philippe, *J Mol Evol* **49**, 496 (Oct, 1999).
31. E. Bapteste, C. Brochier, *Trends Microbiol* **12**, 9 (Jan, 2004).
32. P. Forterre, N. Benachenhou-Lafha, B. Labedan, *Nature* **362**, 795 (Apr 29, 1993).
33. P. Forterre, *Nature* **335**, 305 (1992).
34. P. Forterre, H. Philippe, *Bioessays* **21**, 871 (1999).
35. H. Philippe, P. Forterre, *J. Mol. Evol* **49**, 509 (1999).
36. P. Forterre, H. Philippe, *Biol Bull* **196**, 373 (Jun, 1999).
37. P. Forterre, *ASM News* **63**, 89 (1997).
38. I. B. Rogozin, Y. I. Wolf, A. V. Sorokin, B. G. Mirkin, E. V. Koonin, *Curr Biol* **13**, 1512 (Sep 2, 2003).
39. E. V. Koonin, *J Hered* **100**, 618 (Sep-Oct, 2009).
40. J. C. Sullivan, A. M. Reitzel, J. R. Finnerty, *Genome Inform* **17**, 219 (2006).
41. E. Pennisi, *Science* **317**, 27 (Jul 6, 2007).
42. N. H. Putnam *et al.*, *Science* **317**, 86 (Jul 6, 2007).
43. M. Srivastava *et al.*, *Nature* **454**, 955 (Aug 21, 2008).
44. L. Collins, D. Penny, *Mol Biol Evol* **22**, 1053 (Apr, 2005).
45. E. V. Koonin *et al.*, *Genome Biol* **5**, R7 (2004).
46. V. Anantharaman, E. V. Koonin, L. Aravind, *Nucleic Acids Res* **30**, 1427 (Apr 1, 2002).
47. M. Lynch, A. O. Richardson, *Curr Opin Genet Dev* **12**, 701 (Dec, 2002).
48. C. Rotte, W. Martin, *Nat Cell Biol* **3**, E173 (Aug, 2001).
49. W. Martin, *Proc. R. Soc. Lond. B Biol. Sci* **266**, 1387 (1999).
50. W. Martin, *Curr Opin Microbiol* **8**, 630 (Dec, 2005).
51. T. M. Embley, W. Martin, *Nature* **440**, 623 (Mar 30, 2006).
52. C. G. Kurland, L. J. Collins, D. Penny, *Science* **312**, 1011 (2006).
53. P. Senapathy, *Independent Birth of Organisms*. (Genome Press, Madison, Wisconsin, 1994).
54. R. Regulapati, A. Bhasi, C. K. Singh, P. Senapathy, *PLoS One* **3**, e3456 (2008).
55. P. Senapathy, *Science* **268**, 1366 (Jun 2, 1995).
56. P. Senapathy, *Proc Natl Acad Sci U S A* **85**, 1129 (Feb, 1988).
57. P. Senapathy, *Proc Natl Acad Sci U S A* **83**, 2133 (Apr, 1986).
58. J. U. Bowie, J. F. Reidhaar-Olson, W. A. Lim, R. T. Sauer, *Science* **247**, 1306 (Mar 16, 1990).
59. R. D. Finn *et al.*, *Nucleic Acids Res* **36**, D281 (Jan, 2008).
60. S. Ohno, *Berlin, Springer-Verlag*, (1970).
61. E. Bornberg-Bauer, F. Beaussart, S. K. Kummerfeld, S. A. Teichmann, J. Weiner, 3rd, *Cell Mol Life Sci* **62**, 435 (Feb, 2005).
62. T. Ohta, *Genetica* **118**, 209 (Jul, 2003).
63. F. Hoyle, N. C. Wickramasinghe, *Evolution from Space : A Theory of Cosmic Creationism*. (Simon and Schuster, Inc., New York, 1981).
64. S. W. Roy, W. Gilbert, *Nat Rev Genet* **7**, 211 (Mar, 2006).
65. F. Rodriguez-Trelles, R. Tarrio, F. J. Ayala, *Annu Rev Genet* **40**, 47 (2006).
66. C. A. Orengo *et al.*, *Structure* **5**, 1093 (Aug 15, 1997).
67. M. E. Pettersson, S. Sun, D. I. Andersson, O. G. Berg, *Genetica* **135**, 309 (Apr, 2009).
68. M. Averof, *Curr Opin Genet Dev* **12**, 386 (Aug, 2002).
69. J. Zhang, *TRENDS in Ecology and Evolution* **18**, 292 (2003).
70. S. Bershtein, D. S. Tawfik, *Mol Biol Evol* **25**, 2311 (Nov, 2008).
71. C. G. Kurland, B. Canback, O. G. Berg, *Proc Natl Acad Sci U S A* **100**, 9658 (Aug 19, 2003).
72. A. G. Murzin, *Curr Opin Struct Biol* **8**, 380 (Jun, 1998).
73. D. L. Theobald, D. S. Wuttke, *J Mol Biol* **354**, 722 (Dec 2, 2005).
74. S. Meier, S. Ozbek, *Bioessays* **29**, 1095 (Nov, 2007).
75. G. Vollmer, *Biophilosophie*. (Reclam, Stuttgart, 1995).
76. J. A. Pelesko, *Self Assembly: The Science of Things That Put Themselves Together* (Chapman & Hall/CRC, Boca Raton, 2007).
77. E. Szathmary, F. Jordan, C. Pal, *Science* **292**, 1315 (May 18, 2001).
78. P. Dehal *et al.*, *Science* **298**, 2157 (Dec 13, 2002).
79. E. Pennisi, *Science* **280**, 672 (May 1, 1998).
80. E. Pennisi, *Science* **284**, 1305 (May 21, 1999).
81. E. Sodergren *et al.*, *Science* **314**, 941 (Nov 10, 2006).
82. G. M. Rubin *et al.*, *Science* **287**, 2204 (Mar 24, 2000).
83. E. Pennisi, *Science* **298**, 2111 (Dec 13, 2002).
84. A. Bhasi, P. Philip, V. T. Sreedharan, P. Senapathy, *Genomics* **94**, 48 (Jul, 2009).
85. A. Bhasi, R. V. Pandey, S. P. Utharasamy, P. Senapathy, *Bioinformatics* **23**, 1815 (Jul 15, 2007).
86. C. L. Stewart, K. J. Roux, B. Burke, *Science* **318**, 1408 (Nov 30, 2007).
87. L. J. Terry, E. B. Shows, S. R. Wente, *Science* **318**, 1412 (Nov 30, 2007).
88. N. R. Pace, *Nature* **441**, 289 (May 18, 2006).
89. P. Lopez-Garcia, D. Moreira, *Bioessays* **28**, 525 (May, 2006).
90. A. M. Poole, D. Penny, *Bioessays* **29**, 74 (Jan, 2007).
91. A. Poole, D. Penny, *Nature* **447**, 913 (Jun 21, 2007).
92. P. Senapathy, Molecular Genetics 7, 53 (1988).

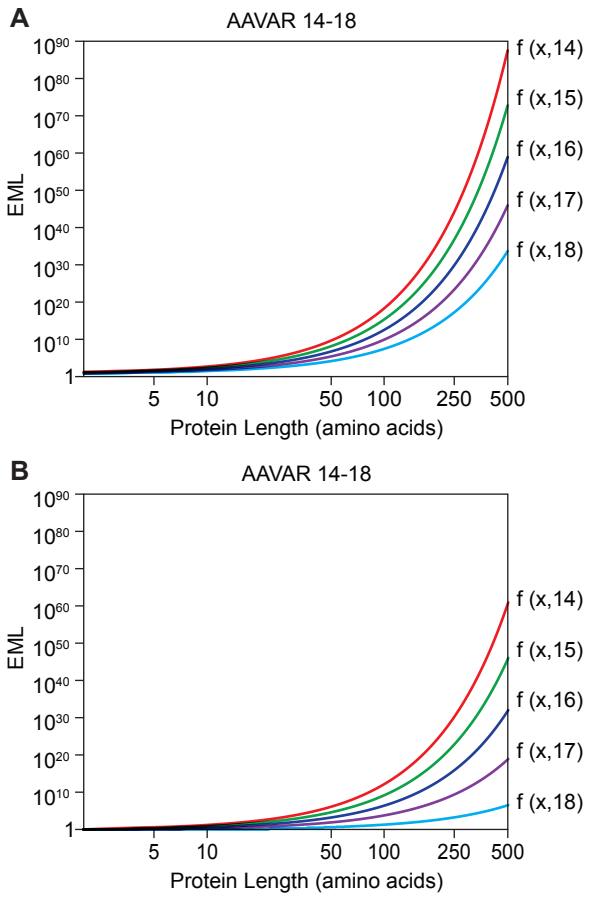
Supplementary Figures and Tables



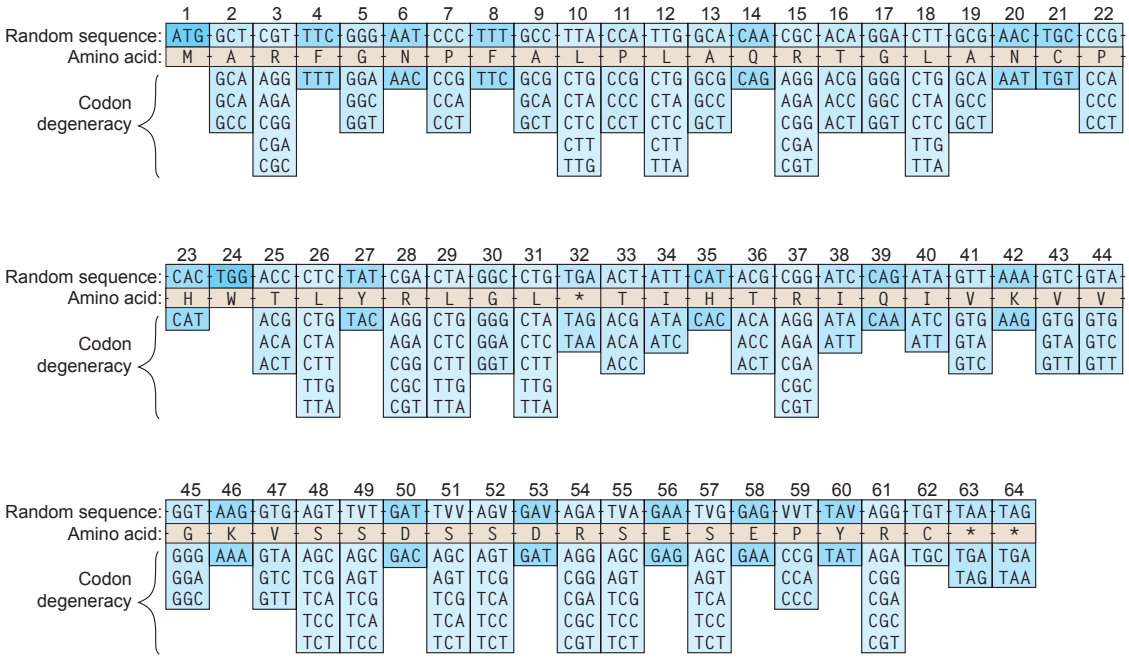
Supplementary Figure 1. | Amino acid variability in a short region of the λ repressor protein. The sequence of the λ repressor in a short region of 17 AA is shown, with the variable AAs above each position. Only position 2, 4 and 10 are invariant. At all other positions, each AA can be changed to any one of the AAs shown above it without altering the structure and activity of the λ repressor.



Supplementary Figure 2. | Stringent restriction of exon lengths in human and random split-genes. The length of the exons from all genes in the human genome, whose coding sequences are >3000 bases and that contain ≥20 exons), as well as whose ORFs were <600 bases, were isolated. Data for five sample genes are shown from a total of 1513 genes that matched these criteria. The length of each exon within a gene was plotted as a horizontal tick mark within a separate vertical bar representing each gene. The lengths of all of the ORFs in each gene on the sense strand were plotted in a separate bar adjacent to each exon bar. A) Data for five human genes, whose Gene IDs are shown on X-axis below each bar group. B) Data from five random split genes predicted by GenScan software (See Methods).



Supplementary Figure 3. | **Effect of increasing the length of protein-coding sequence on its EML.** The probability of the coding sequence for a protein sequence with AAVAR (PFAM database), and the EML for the occurrence of the coding sequence for any one of its variable AA sequences in a random DNA, were computed as follows: An average degeneracy of 3.05 codons (A) and 3.45 codons (B) for each AA, and an average AAVAR of 14, 15, 16, 17 or 18 AAs per AA sequence position, were used in the AVARIA algorithm (see Methods). The EMLs of the random DNA sequence for the occurrence of the coding sequence for proteins with varying sequence lengths (1-500 AAs), and for different average AAVARs (14, 15, 16, 17 and 18 AAs per sequence position) were plotted.



Supplementary Figure 4. | A higher average codon degeneracy per amino acid due to higher frequency of more degenerate codons in a random DNA sequence. The frequency of degenerate codons is higher than the non-degenerate codons in a random DNA sequence. This increases the codon degeneracy from 3.05 (61/20) codons per codon position to about 3.85 (235/61) codons per codon position. The standard one letter codes for amino acids are shown.

Supplementary Table 1. | **Domains from the PFAM database used for constructing model proteins.** Twenty-six unique protein domains, each with a different structure and function, were selected from the PFAM database (see Supplementary Table 2 & Methods). Each was assigned a one-letter code, different combinations of which were used to construct the model proteins shown in Table 2. The average AAVAR in a given protein domain was computed by dividing the sum of the AAVARs that occur at all of the sequence positions of that domain in the PFAM database by the length of the domain sequence.

Domain Code	Domain Name	PFAM ID	Domain Length (AA)	Mean AAVAR (AAs per sequence position)	Probability	EML (bases)
A	Homeobox*	PF00046	55	11.6	5.11E-18	3.23E+19
B	Ankyrin repeat	PF00023	16	19.1	1.83E-01	2.62E+02
C	HEAT repeat	PF02985	13	18.7	2.15E-01	1.81E+02
D	F-box domain	PF00646	31	18.2	8.78E-03	1.06E+04
E	PPR repeat	PF01535	33	18.0	7.22E-03	1.37E+04
F	Histidine kinase-, DNA gyrase B-, and HSP90-like ATPase	PF02518	18	17.9	3.47E-02	1.56E+03
G	CBS domain pair	PF00571	67	17.8	5.27E-05	3.81E+06
H	Cyclic nucleotide-binding domain	PF00027	67	17.4	5.85E-06	3.44E+07
I	Glycosyl transferase family 2	PF00535	111	17.3	4.37E-10	7.62E+11
J	AMP-binding enzyme	PF00501	233	17.2	4.67E-20	1.50E+22
K	Metallo-beta-lactamase superfamily	PF00753	79	17.2	6.14E-09	3.86E+10
L	Two component regulator propeller	PF07494	16	17.1	2.81E-02	1.71E+03
M	haloacid dehalogenase-like hydrolase	PF00702	123	17.1	1.63E-11	2.26E+13
N	LysR substrate binding domain	PF03466	97	17.1	4.18E-09	6.96E+10
O	Threonine leader peptide	PF08544	31	16.9	1.50E-03	6.22E+04
P	GAF domain	PF01590	38	16.9	4.73E-04	2.41E+05
Q	TonB dependent receptor	PF00593	73	16.8	2.71E-07	8.09E+08
R	Tetratricopeptide repeat	PF07719	32	16.8	4.16E-04	4.30E+07
S	Universal stress protein family	PF00582	60	16.8	4.19E-06	2.31E+05
T	F-box associated	PF07735	17	16.7	2.74E-02	1.86E+03
U	OB-fold nucleic acid binding domain	PF01336	39	16.7	1.07E-04	1.10E+06
V	Xylose isomerase-like TIM barrel	PF01261	61	16.6	7.77E-07	2.36E+08
W	Integrase core domain	PF00665	63	16.6	6.82E-08	2.77E+09
X	EF hand	PF00036	12	16.5	5.43E-02	6.63E+02
Y	Phosphotransferase enzyme family	PF01636	57	16.4	5.55E-08	3.08E+09
Z	Leucine Rich Repeat	PF00560	9	16.0	5.72E-02	4.72E+02
Collagen	Collagen*	PF01391	60	7.5	6.91E-33	2.60E+34

* Though these domains have a low AAVAR in Pfam, we used them to demonstrate that the split genes coding for these proteins that are supposed to be highly evolved (e.g. the repetitive pattern of collagen, Gly-X-Pro) also occurs in random DNA at essentially the same probability as that of a protein with a unique sequence.

Supplementary Table 2. | **The probability and the EML of the contiguous coding DNA sequence for the protein domains from the PFAM database in a random DNA sequence.** The average AAVAR of each of the 9318 domains available in the PFAM database were computed (see Methods). The domain with the highest AAVAR for each domain length represented in the PFAM database was identified, and a sample of 30 domains from a total of 666 domain lengths are shown. The probability and the EML for the occurrence of the coding sequence for each of these domains in a random DNA sequence were computed using the AVARIA program based on the AAVAR and CD at each AA position within a given domain.

Protein name	PFAM ID	Sum of AA variability	Domain length (AA)	Mean AA variability	Probability	EML (bases)
WD domain, G-beta repeat	PF00400	487	25	19.5	1.73E-01	4.33E+02
Ankyrin repeat	PF00023	306	16	19.1	1.83E-01	2.62E+02
HEAT repeat	PF02985	243	13	18.7	2.15E-01	1.81E+02
Radical SAM superfamily	PF04055	1147	63	18.2	3.22E-05	5.88E+06
F-box domain	PF00646	564	31	18.2	8.78E-03	1.06E+04
PPR repeat	PF01535	595	33	18.0	7.22E-03	1.37E+04
Histidine kinase-, DNA gyrase B-, and HSP90-like ATPase	PF02518	322	18	17.9	3.47E-02	1.56E+03
CBS domain pair	PF00571	1192	67	17.8	5.27E-05	3.81E+06
Cyclic nucleotide-binding domain	PF00027	1166	67	17.4	5.85E-06	3.44E+07
Glycosyl transferase family 2	PF00535	1921	111	17.3	4.37E-10	7.62E+11
AMP-binding enzyme	PF00501	4017	233	17.2	4.67E-20	1.50E+22
Metallo-beta-lactamase superfamily	PF00753	1359	79	17.2	6.14E-09	3.86E+10
Two component regulator propeller	PF07494	274	16	17.1	2.81E-02	1.71E+03
haloacid dehalogenase-like hydrolase	PF00702	2099	123	17.1	1.63E-11	2.26E+13
LysR substrate binding domain	PF03466	1654	97	17.1	4.18E-09	6.96E+10
MORN repeat variant	PF07661	272	16	17.0	4.60E-02	1.04E+03
Threonine leader peptide	PF08544	524	31	16.9	1.50E-03	6.22E+04
GAF domain	PF01590	641	38	16.9	4.73E-04	2.41E+05
TonB dependent receptor	PF00593	1228	73	16.8	2.71E-07	8.09E+08
Tetratricopeptide repeat	PF07719	536	32	16.8	4.16E-04	4.30E+07
Universal stress protein family	PF00582	1005	60	16.8	4.19E-06	2.31E+05
F-box associated	PF07735	284	17	16.7	2.74E-02	1.86E+03
OB-fold nucleic acid binding domain	PF01336	650	39	16.7	1.07E-04	1.10E+06
Xylose isomerase-like TIM barrel	PF01261	1015	61	16.6	7.77E-07	2.36E+08
Integrase core domain	PF00665	1043	63	16.6	6.82E-08	2.77E+09
EF hand	PF00036	198	12	16.5	5.43E-02	6.63E+02
Peptidase M16 inactive domain	PF05193	297	18	16.5	3.46E-02	1.17E+06
Acetyltransferase (GNAT) family	PF00583	627	38	16.5	9.75E-05	1.56E+03
Tetratricopeptide repeat	PF00515	544	33	16.5	1.56E-04	6.33E+05
Phosphotransferase enzyme family	PF01636	936	57	16.4	5.55E-08	3.08E+09