54. A. Tomlinson, D. D. L. Bowtell., E. Hafen, G. M. Rubin, *Cell* **51**, 143 (1987).
55. Polyclonal antisera specific for *sca* products were obtained after immunizing BALB/c mice with a TrpE-Sca fusion protein [T. J. Koener, J. E. Hill, A. M. Myers, A. Tzagoloff, *Methods Enzymol.*, in press] lacking only the NH_2-terminal 19 amino acids of the putative *sca* signal peptide.
56. pcs6 DNA was sequenced on both strands with the shotgun cloning method and chain-terminating inhibitors [A. T. Bankier and B. G. Barrell, in *Techniques in the Life Sciences, B5, Nucleic Acid Biochemistry* (B508), 1-34, R. A. Flavell, Ed. (Elsevier, New York, 1983)], and the modified T7 polymerase "Sequenase" [S. Tabor and C. C. Richardson, *Proc. Natl. Acad. Sci. U.S.A.* **84**, 4767 (1987)]. Oligonucleotides derived from the cDNA sequence and its complement were used as primers to sequence corresponding regions of a double-stranded genomic DNA template. Introns were identified from the discontinuity of the cDNA and genomic sequences. The intron-exon structure obtained was consistent with the cDNA and genomic restriction maps, with the distribution of genomic restriction fragments complementary to cDNA in hybridization experiments (see *19* for the genomic organization of the *sca* exons), and with the consensus sequences for splice donor and acceptor sites [S. M. Mount, *Nucleic Acids Res.* **10**, 459 (1982)]. Mutant DNA's were obtained by means of the polymerase chain reaction. Genomic DNA from *mutant/deficiency* or from wild-type strains was amplified with oligonucleotide primers corresponding to sequences flanking the four exons. The double-stranded reaction products were either sequenced directly with oligonucleotide primers, or cloned into m13 and sequenced from single-stranded recombinant bacteriophage DNA. Sequence differences were confirmed by sequencing products from more than one independent polymerase chain reaction.
57. G. Von Heijne, *Nucleic Acids Res.* **14**, 4683 (1986).
58. At least one other *Drosophila* protein, the product of the *wingless* gene, is known to be secreted despite lacking the n-region [F. R. Rijsewijk *et al.*, *Cell* **50**, 649 (1987); M. Van den Heuvel, R. Nusse, P. Johnston, P. A. Lawrence, *ibid.* **59**, 739 (1989)].
59. M. Hortsch, A. J. Bieber, N. H. Patel, C. S. Goodman, *Neuron* **4**, 697 (1990).
60. Supported in part by postdoctoral fellowships from the Damon Runyon–Walter Winchell Cancer Research Fund (N.E.B.) and EMBO (M.M.). We thank the laboratories of S. Artavanis-Tsakonas, J. Campos-Ortega, and L. Y. and H. N. Jan for *Drosophila* stocks and unpublished information, M. Simon for the negative printed in Fig. 2A, E. Bier, R. Cagan, R. Doolittle, C. Goodman, and P. Hoppe for interesting discussions, and R. Carthew, E. Ferguson, J. Fischer, M. Freeman, C. Goodman for comments on the manuscript. Sequences are available from GenBank under the accession number M37703.

10 July 1990; accepted 26 October 1990

How Big Is the Universe of Exons?

ROBERT L. DORIT, LLOYD SCHOENBACH, WALTER GILBERT

If genes have been assembled from exon subunits, the frequency with which exons are reused leads to an estimate of the size of the underlying exon universe. An exon database was constructed from available protein sequences, and homologous exons were identified on the basis of amino acid identity; statistically significant matches were determined by Monte Carlo methods. It is estimated that only 1000 to 7000 exons were needed to construct all proteins.

MOST GENES IN COMPLEX EUKARYOTES CONSIST OF short exons separated by long introns. In one view, genes are assembled, via intron-mediated recombination, from exon modules that code for functional domains, folding regions, or structural elements (*1*, *2*). Such models portray introns as a retained primitive feature. Alternatively, the phylogenetic distribution of introns has led to arguments that introns are a derived feature of eukaryotic genomes, the result of bursts of parasitic elements invading early (and continuous) eukaryotic coding regions (*3*, *4*).

The hypothesis of exon shuffling proposes that complex genetic information is built up by joining previously independent exons, thus giving rise to more complex proteins and to novel enzymatic functions. This view of the modular assembly of extant genes is supported by the common structural features of certain large gene superfamilies, such as the immunoglobulin-like superfamily (*5*), and by the examples of exon reuse observed in the mosaic structure of the LDL (low density lipoprotein) receptor and the EGF (epithelial growth factor) precursor (*6*). In other gene superfamilies, the older intron-exon gene structure is still apparent in certain representatives, while other members of the same family have lost introns (possibly through retroposition of a mature message) to produce genes with longer and more complicated exons, but with few or no remaining introns. An example of this pattern is the opsin superfamily, which includes genes with four introns as well as genes for beta-adrenergic receptors, which have no introns at all (*7*).

The ancient character of introns is also supported by data suggesting that introns antedate the divergence of plants and animals a billion years ago (*8*). Intron-exon structures may also predate the endosymbiotic incorporation of chloroplasts and mitochondria, which occurred about 2 billion years ago (*9*). Introns may, in fact, antedate the first branchings of life on Earth: the first protogenes may have already displayed intron-exon structure. The original exons may have been 15 to 20 amino acids long; processes of intron sliding and intron loss leading to more complex exons have produced the present day spectrum (*2*).

In this article, the frequency of exon shuffling events is surveyed in order to address the following question: How many different exons were required to generate the current protein diversity? We have identified homologous exons (those of common evolutionary origin) on the basis of amino acid sequence similarity. To the extent that every exon in an underlying universe of exons has an equal probability of being incorporated into a gene, we can then estimate the size of that underlying universe by determining how frequently homologous exons appear in nonhomologous genes.

We first constructed a database of all known exons. The available databases contain large numbers of homologous gene sequences; we eliminated such duplication in order to obtain a collection of exons derived solely from independent genes, unrelated by direct descent. We then made pairwise comparisons of all these independent exons to identify statistically significant sequence similarities, which, we argue, indicate exon homology.

Finally, using a simple sampling model, we took this number of exon repeats to estimate the size of the exon universe. If we survey n exons that have been drawn with replacement from an underlying set of size N, we expect the number of repeats to be given by the product of $n(n - 1)/2$, the number of pairs of objects in the collection, and the probability that any pair will match, $1/N$. The number of single repeats is thus $n(n - 1)/2N$. Accordingly the expectation for triple repeats is $n(n - 1)(n - 2)/6N^2$ and so forth (*10*).

The authors are with the Department of Cellular and Developmental Biology, Harvard University, 16 Divinity Avenue, Cambridge, MA 02138.

The exon database. The database of exons was drawn from the eukaryotic genes of known structure recorded in the GenBank and EMBL computer databases (*11*). Using the DNA sequence and the features table, we wrote computer programs that translated each exon. We then inspected the resulting collection of exons and corrected by hand those cases in which the exon boundary had been incorrectly specified (by typographical error). To obtain a distilled database containing exons drawn only from putatively nonhomologous proteins, we first purged the database of homologous genes from different species, retaining a representative human sequence wherever possible. We then removed closely related or duplicated genes, such as the multiple globin sequences (alpha, beta, embryonic, and myoglobin), again retaining but a single example, and culled repeating structures within single genes (such as multiple repeats of a single exon that make up the collagen gene, and the triply repeated domains in serum albumin and ovomucoid). Recurrent elements in a gene superfamily, such as the multiple repeats of the immunoglobulin fold in the immunoglobulin superfamily, or the multiple occurrences of the serine-protease domains in the family that includes the blood-clotting factors, were also pared down to single representative examples. Our initial exon sequence comparisons still included occasional exon pairs displaying more than 80 percent amino acid sequence similarity—one member of each such pair was discarded. Finally, we arbitrarily excluded all exons shorter than 20 amino acids, both because of the high sequence similarity that would be required to establish statistical significance and because this size class contains many signal sequences, which display unusually biased amino acid compositions. These extensive refinings eventually reduced the original database to less than half its size, leaving us with a purged collection containing 1255 exons. The final distribution of exon lengths peaks around 40 to 50 amino acids (Fig. 1).

Criteria for exon similarity and statistical significance. We compared the amino acid sequences of individual exons by making pairwise comparisons, scoring only exact amino-acid matches, and allowing no gaps. A given exon of length N (number of amino acid residues) was compared to all exons of length N to $N+10$, thus allowing for small variations in exon length resulting from insertions, deletions, or splice-site shifts. To optimize alignments, we allowed exons to slide up to five amino acids out of end register in either direction during each pairwise comparison and recorded the best percentage match (the number of matching amino acids times 100, divided by the length of the shorter exon). For computational convenience, the exon database was arbitrarily divided into nine length classes (20 to 29, 30 to 39, and so on). Inset **a** of Fig. 4 shows the output of a representative similarity run for exons of length 40 to 49. The number of events recorded in the histogram corresponds to the number of exon pairs compared in the search. The histogram of similarity values (for exons of length 40 to 49 a.a.) is normally distributed about a mean similarity of 12 percent (with a variance of 8 percent). Roughly speaking, if all amino acids appeared at the same frequency, one expects an average match of 5 percent, improved about three standard deviations by the sliding algorithm. Our study involved a total of 215,166 pairwise exon comparisons, and about 3 million actual comparisons.

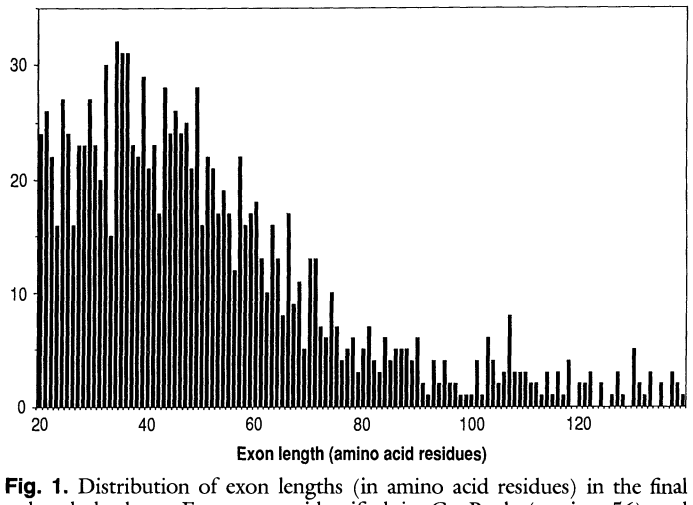
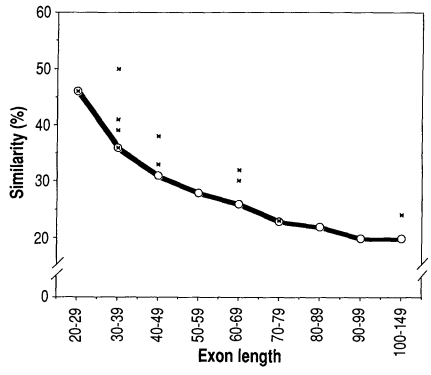


Fig. 1. Distribution of exon lengths (in amino acid residues) in the final reduced database. Exons were identified in GenBank (version 56) and EMBL (version 15). That collection was purged of repeats, homologous genes, and superfamily relationships by repeated rounds of analysis. Exons shorter than 20 amino acids were excluded from the analysis.

Fig. 2. Significance cutoffs as a function of exon length. The cutoff is defined as the highest value of the similarity statistic that occurs only once in 20 simulation runs. Asterisks indicate sequence similarity for each identified case of exon shuffling.



We developed criteria for the statistical significance of exon sequence matches by carrying out repeated Monte Carlo simulations, each time randomizing the sequence of every exon and comparing that randomized exon with the original data set. Each random sequence was constructed by sampling (with replacement) from an amino-acid pool derived from all the exons in the real data set. Thus, the amino acid compositions of the randomized and real data sets are identical, but a biased composition of an actual real exon is not likely to reappear in the randomized data set. The comparison program run on these randomized exons establishes the level of exon similarities expected by chance alone. By carrying out 20 different randomized runs, we determine for each range of exon sizes the highest similarity value that occurred only once in 20 runs, and we take this as a cutoff value (Fig. 2). Any match between two real exons that is greater or equal to the cutoff value will be statistically significant, since it is likely to occur by chance no more than once in 20 trials. This similarity cutoff is quite stringent; in order to be considered homologous, exons must display sequence similarities ranging from 46 percent identity for exons of length 20 to 29 down to about 20 percent identity for exons of length 100.

Fourteen exon pairs exhibited amino acid similarity greater or equal to the required cutoff values (Table 1). The similarity values of these matches, relative to the simulation cutoffs, appear in Fig. 2, and the matching pairs themselves are listed in Table 2. Some of these matches have been recognized before (shown by an asterisk [*]), while others are new. The known examples include the collagen-like domain of mannose-binding protein (*12*); the EGF-like domains documented in both Factor IX (*13*) and Factor XII (*14*); the collagen motif characteristic of the β-chain of complement C1q (*15*), and the thyroglobulin-like alternatively spliced exon (*6*) of the Ia antigen–associated chain (*16*). The examples of exon shuffling (Table 1) illustrate a number of themes. A motif, encoded by a single exon, may be performing a similar function in two otherwise unrelated proteins. For example, the first exons of collagenase and major urinary protein serve (at least in part) as signal peptides in both proteins (*17*, *18*). Exon 4 of the chloroplast *psbA* gene and exon 17 of band 3 protein function as membrane-spanning domains

(*19*, *20*). In contrast, exon 2 of β-lymphotoxin (*21*) and exon 3 of the asialoglycoprotein receptor (*12*, *22*) represent a single hydrophobic domain playing different roles—as a signal sequence in the first protein and as a transmembrane segment in the latter. Table 1 contains several exons derived from collagens, intermediate filaments, or other structural proteins. This pattern may reflect the limited number of basic motifs that can serve as connective or matrix proteins, as well as the evolutionarily conservative character of such protein sequences.

To verify that these matches represented genuine cases of exon shuffling, we compared the proteins from which the exons were derived in their entirety, seeking to maximize the alignment across the whole protein by allowing gaps (*23*). In all the cases that we describe, the amino acid similarity across the entire protein, or across any region (excluding the exon pair we identify) is significantly lower than that of our matched exon pair. We present two examples of exon matches in the context of their proteins (Fig. 3). In Fig. 3A the intron positions are in similar phase, clearly the surrounding sequences and exons of these genes are not related to the degree exhibited by the relevant exon sequences. In Fig. 3B the intron junctions have drifted in both position and phase.

Table 1. Identified cases of exon homology. The identity, length, and sequence similarity of exon pairs are shown, arranged by decreasing similarity. Asterisks indicate previously identified exon homologies.

Protein	Exon	Exon sizes (a.a. residues)	Similarity (%)
Human α-1 (II) collagen (*32*)	[X]	36	50*
Rat mannose-binding protein A (*12*)	[2]	38	
Human apolipoprotein B-100 (*33*)	[1]	24	46
Human EGF receptor (*34*)	[1]	29	
Human blood coagulation factor XII (*14*)	[7]	34	41*
Human factor IX gene (*13*)	[4]	37	
Human pro-α-1 type I collagen (*35*)	[47]	34	38
Human elastin (*36*)	[8]	41	
Mouse major urinary protein (*18*)	[1]	32	38
Rabbit collagenase (*17*)	[1]	34	
Chicken steroid inducible hsp (*37*)	[7]	40	38
Human neurofilament subunit NF-L (*38*)	[4]	47	
Human lymphotoxin (TNF-β) (*21*)	[2]	33	36
Rat asialoglycoprotein receptor (*22*)	[3]	38	
Schizophyllum 1G2 gene (fruiting) (*39*)	[1]	40	33
Human fibronectin (*40*)	[1]	49	
Chicken fps proto-oncogene (*41*)	[8]	40	33
Human neurofilament subunit NF-L (*38*)	[4]	47	
Mouse α-2 type IV collagen (*42*)	[5]	60	32*
Human complement C1q B-chain (*15*)	[1]	64	
Murine Ii gene, Ia antigen–associated (*16*)	[6b]	63	30*
Bovine thyroglobulin (*43*)	[18]	64	
Silkmoth chorion (*44*)	[2]	108	24
Mouse keratin, intermediate filament (*45*)	[7]	112	
C. reinhardtii chloroplast *psbA* gene (*19*)	[4]	77	23
Mouse band 3 (*20*)	[17]	84	
Human serum albumin (*46*)	[4]	70	23
Human K6b epidermal keratin (*47*)	[7]	73	

Table 2. Comparison of pairwise similarity values for the real and scrambled Monte Carlo simulations in the uppermost 5 percent of the distributions. The table displays exon length intervals, numbers of actual comparisons, the similarity value that specified the top 5 percent cutoff, the number of matches for both the Monte Carlo and the actual runs above the 5 percent cutoff, the excess of the real matches relative to the simulations, and Fisher's exact *P* value for that excess (one degree of freedom). The Monte Carlo value is the mean of 20 simulations.

Exon lengths	Comparisons (no.)	Cutoff (percent)	Monte Carlo	Real	Excess matches	*P*
20–20	53251	21	3801	4072	271	0.000
30–39	61144	19	3359	3625	266	0.001
40–49	54190	17	3557	3775	218	0.004
50–59	27191	16	1614	1633	19	0.372
60–69	11678	15	839	914	75	0.033
70–79	3693	14	336	372	36	0.084
80–89	1778	14	91	103	12	0.23
90–99	558	13	56	54	−2	0.46
100–149	1683	13	103	102	−1	0.50

These 14 exon matches predict an underlying exon universe of 56,000 sequences. Because we rely on amino acid sequence identity in our analysis, allow no gaps in the alignment process, and demand such a high degree of similarity for significance, we are likely to underestimate the number of homologous exon pairs, and hence overestimate the universe. Certain standard examples of exon shuffling, such as the LDL receptor, were missing from the database and are so are not in our table. We have also omitted certain well-known cases of exon shuffling, such as the serine protease domain, a conspicuous feature of a large family of proteins (*24*–*26*), and the various immunoglobulin motifs shared by the members of the immunoglobulin (Ig) superfamily (*5*). Both the serine protease and immunoglobulin domains span more than a single exon and thus do not meet the specific criteria of this study.

Cases where intron loss leads to the incorporation of a shuffled exon into a larger protein domain are also likely to be missed given the limited size of our search window (±10 amino acids). Finally we demand 30 to 40 percent sequence identity for most comparisons. Proteins (or protein domains) that have drifted very far in amino acid sequence may nonetheless retain their three-dimensional similarity: one can identify structural homologies in circumstances where only 10 percent of the amino acid sequence is conserved (*27*). Thus, many exons with common evolutionary origins will not be recognizable by amino acid sequence similarity alone. We believe that this calculation of 56,000 members could easily be a five- to tenfold overestimate of the size of the exon universe.

The wedge calculation. In examining the distribution of similarity values, we noticed an excess of real matches (relative to our simulations) at the high end of the similarity distribution. Is this excess statistically significant? In the earlier Monte Carlo calculations, the randomized sequences were chosen to match the overall amino acid composition of the exon database. Two real exons that share a highly skewed amino acid composition then would likely match above the significance criterion. In that first calculation, we considered such a match to be evidence of evolutionary homology. To test more stringently for any excess of real matches, we carried out new Monte Carlo simulations, this time scrambling the amino acid sequence of each exon (creating anagrams of the real exons) and hence preserving the compositional bias of each particular exon. We averaged 20 simulations to produce a baseline distribution against which to compare the real exon similarities. The data for exon comparisons in the 40 to 49 window are shown in Fig. 4. The full

curve in the inset shows the distribution of the matches in the real data, and the enlargements show, for the right-hand tail of the similarity value distribution, the differences between the matches found with real data and those from scrambled exons. To estimate the significance of this excess, we took the top 5 percent of the distribution of the similarity statistic, compared the real and simulation distributions, and determined the significance of the excess (Fisher's exact test). In this top 5 percent includes all matches above 17 percent sequence similarity; the excess of real over scrambled is shown as the black tips of the bars. Table 2 shows the comparisons of the top 5 percent of the distribution for each exon size class, which we refer to as "the wedge." Significant excesses of real matches do exist for most of the exon size classes (*28*). The total excess sums to 830 matches over all significant intervals. This number of matches, arising in the sample of 1255 exons, predicts an underlying exon universe of just 950 exons.

This low number for a universe of fundamental shapes suggests that our database of 1255 exons includes examples of most of the original exon universe. The calculation demonstrates that the number of matches is significantly above the expectation based on stochastic sequence similarity, even after discounting shared compositional biases. (For example, two exons that consist of 50 percent leucines will tend to match for that reason alone; this calculation excludes such matches.) While this approach provides substantial statistical power, it does not identify specific pairs of homologous exons. One cannot deduce which matches in the wedge are biologically meaningful and which constitute the random background.

A further test also shows that the excess of matches in the wedge is likely due to exon shuffling. One might have argued that the excess of amino acid identities in the real sequences reflects some convergent or recurrent theme of protein structure, apparent in real sequences but absent in scrambled sequences. For example, some hidden sequence regularity in α helices or some correlation in dipeptide or tripeptide frequencies could cause protein sequences to match against each other at a frequency above random expectation. To test for a strong effect of such features, we carried out further Monte Carlo simulations, this time creating pseudo-exons by transposing a block of sequence from the front (NH_2-terminus) to the rear (COOH-terminus) of an exon and then comparing this pseudo-exon against the other members of the database. This rearrangement of sequence blocks within an exon would preserve sequence similarity due to local features but destroy any similarity that depended on the actual boundaries of the exon. Simulations transposing blocks of 15 to 25 amino acids for the 40 to 49 window gave results that agree with those of the scrambled exon simulation; there is the same significant excess of matches of real exons over both the "scrambled exon" simulations and the "block-transposed exon" simulations. The similarity between real exons thus does not derive from small stretches of local identity but depends instead on the position of the outer boundaries. We expect this outcome if the excess of matches displaying high scores comes about because of true exon homology, where the protein sequence within the exons has been preserved with respect to the positions of the introns.

Reliability of the estimates. We present two different methods to estimate the size of the exon universe. The first identifies 14 examples of exon homology and estimates an underlying exon universe of about 56,000 members. The second identifies a significant total excess of high pairwise similarities (in the top 5 percent of the distribution) corresponding to 830 cases of exon shuffling, thus reflecting an underlying universe of 950 exons. We believe that the first calculation overestimates the size of the exon universe, while the second calculation, although reflecting some significant aspect of exon structure, may be an underestimate. Our best expectation lies somewhere in between. The geometric mean of these numbers is about 7000. Our final expectation, on balance, is between 1000 and 7000 for the size of the exon universe.

Our conclusion may be exaggerated. We may not have succeeded in eliminating all of the biases inherent in the computer databases. The sequence similarity we observe might be the result of convergent evolution, although there is no a priori reason to suspect that any convergence would respect exon boundaries. Furthermore, this search could examine only eukaryotic sequences. If the prokaryotic proteins turn out not to be related to the exon peptide patterns apparent in the eukaryotic sequences, the number of total patterns in the universe would necessarily increase. Finally, the traditional

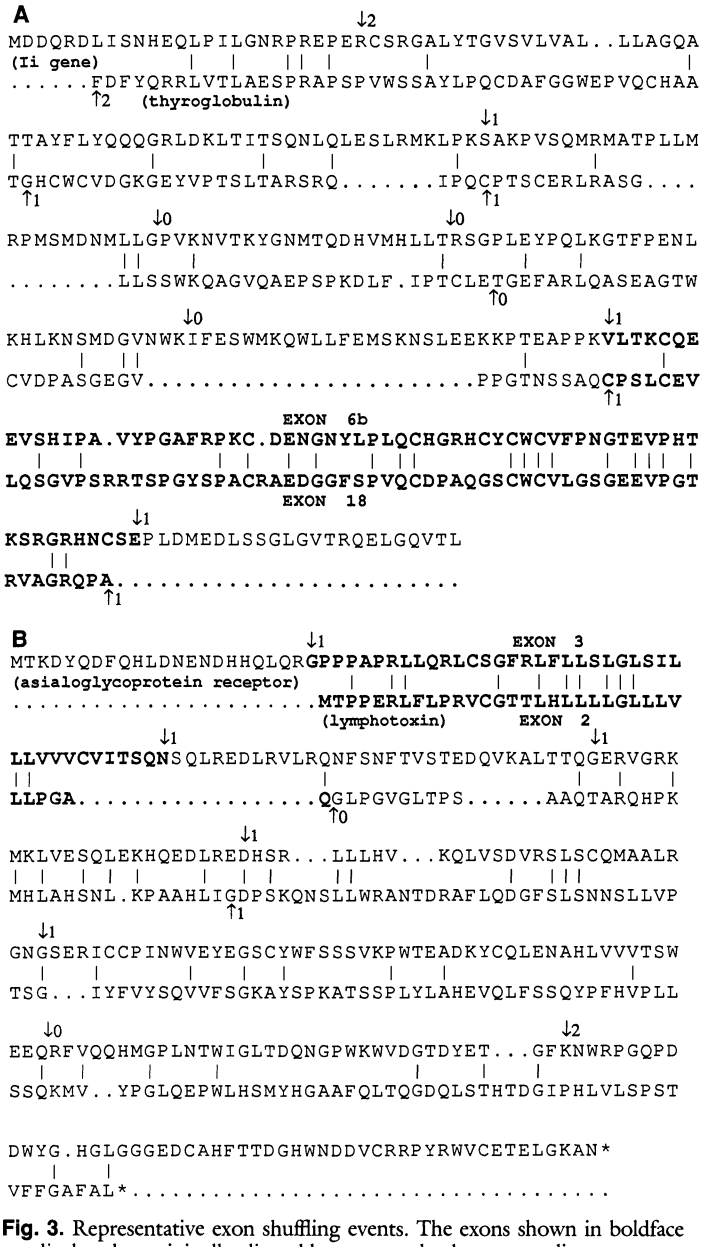


Fig. 3. Representative exon shuffling events. The exons shown in boldface are displayed as originally aligned by our search; the surrounding sequences and exons are simply displayed for contrast. Vertical arrows indicate the phase of the intron/exon boundaries. (**A**) Comparison between the alternatively spliced exon (6b) of the murine Ii gene (*16*) and exon 18 of bovine thyroglobulin (*43*). (**B**) Comparison of exon 3 of the rat asialoglycoprotein receptor (*22*) and exon 2 of human lymphotoxin (*21*). The surrounding protein sequences are aligned to maximize overall sequence similarity. Abbreviations for the amino acid residues are: A, Ala; C, Cys; D, Asp; E, Glu; F, Phe; G, Gly; H, His; I, Ile; K, Lys; L, Leu; M, Met; N, Asn; P, Pro; Q, Gln; R, Arg; S, Ser; T, Thr; V, Val; W, Trp; and Y, Tyr.

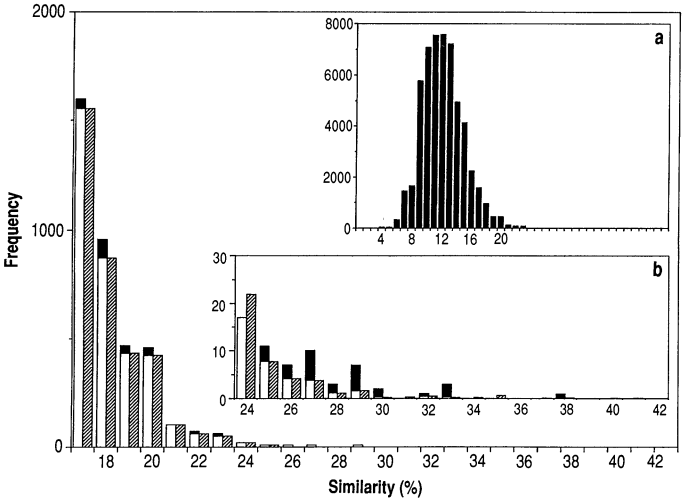


Fig. 4. Representative distribution of the pairwise similarity values for real and randomized exons of length 40 to 49. The figure shows the highest 5 percent of the distribution for the real and simulation comparisons: cross-hatched bars, simulation results; white bars, real exon comparisons. Excess of real matches (wedge) is displayed as black boxes. Inset **a** shows the total distribution of similarity scores for real exon pairwise comparisons. Inset **b** enlarges the rightmost tail of the distribution.

strategies of molecular biology may constrain the kinds of sequences found; systematic whole genome sequences may reveal novel classes of proteins and exons.

The surprisingly small size of our estimate emphasizes the finite character of the underlying exon universe. The number of possible 40-amino acid–long structures, 20^{40} or 10^{52}, is a much larger domain of shapes than the 10^3 to 10^4 that we here predict. Although rules restricting the folding of amino acid chains may have eliminated a large number of amino acid sequences, chance alone may account for which specific elements were in the initial set of exons that gave rise to modern proteins. With a sufficient set of three-dimensional shapes, stabilities, and rudimentary functions, the evolution of proteins could be set in motion.

Several recent studies on protein structure also suggest that the number of possible three-dimensional shapes is quite small. Jones and his co-workers observed that if one connects two points with a loop of 6 or 7 amino acids, only a limited number of Cα patterns will fit (*29*). Unger and co-workers (*30*) have recently shown, by examining the three-dimensional Cα structures of each set of six amino acids in the crystallographic database, that the structures of all hexamers can be clustered into only 80 types rather than 10^8. These observations suggest that the range of shapes in proteins is not as extensive as one might have feared. Recently, Sander and Schneider (*31*) have sought to establish the extent of amino acid sequence similarity that predicts structural (three-dimensional) "homology" between two protein sequences. Their analysis determines a curve of threshold similarity parallel, but slightly more stringent, than our cutoff curve (Fig. 2) and strongly supports the argument that our homologous exon pairs may indeed adopt similar three-dimensional configurations within the different proteins.

Our argument also helps to elucidate the processes of protein evolution. The complexity of modern proteins would have been generated by simply combinatorial arrangements of a limited number of units of structure and function. Particular functional units—DNA-binding motifs, for example, or metal-binding domains—reappear in different contexts to confer new functions on novel exon combinations.

The consequences of a combinatorial search through sequence space are profound. In contrast to a random amino-acid search, the modular building of proteins entails a faster but far more restricted exploration of possible solutions. A 200 amino acid protein may result from a linear search through only 25,000 possible combinations (five modules of 40 amino acids each; 5000 possible exon shapes), rather than the 20^{200} solutions that comprise a full amino-acid-by-amino-acid search.

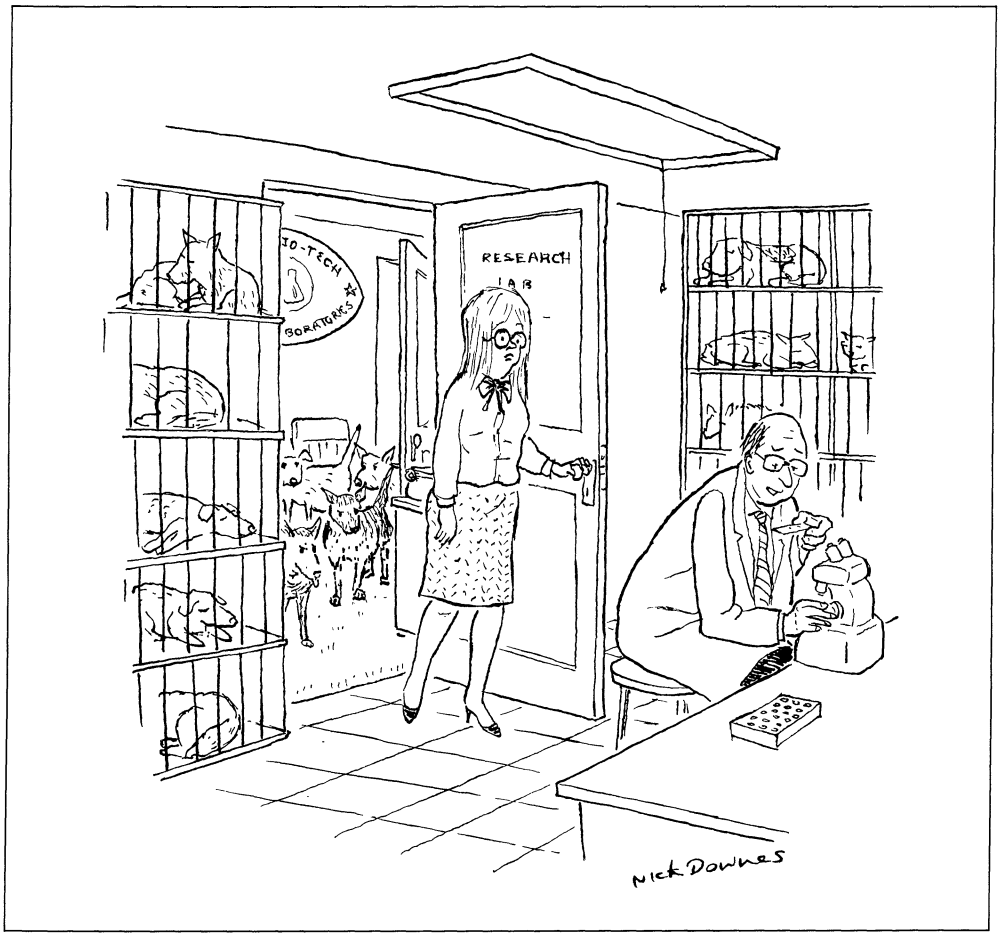
History constrains all evolutionary phenomena. We have argued that modern protein diversity represents only a very limited exploration of sequence space, an exploration constrained by the success of earlier motifs. While we could argue that the corner of sequence space occupied by modern proteins represents the best of all possible worlds, a selective optimum reached after a careful evolutionary walk through all of sequence space, this seems extremely unlikely. The processes that result in protein diversification—exon reassortment initially, followed by gene duplication and divergence—sharply limit protein sequence diversity. Extant proteins may well lie at local, not global, optima.

REFERENCES AND NOTES

1. W. Gilbert, *Nature* **271**, 501 (1978); W. F. Doolittle, *ibid.* **272**, 581 (1978); C. C. F. Blake, *ibid.* **306**, 535 (1983); C. C. F. Blake, *ibid.* **277**, 598 (1979).
2. W. Gilbert, *Cold Spring Harbor Symp. Quant. Biol.*, **52**, 901 (1987).
3. T. Cavalier-Smith, *Nature* **315**, 283 (1985); J. Rogers, *ibid.*, p. 458.
4. D. A. Hickey, B. F. Benkez, S. M. Abukashawa, *J. Theor. Biol.* **137**, 41 (1989); D. A. Hickey and B. F. Benkez, *ibid.* **121**, 283 (1986).
5. T. Hunkapiller and L. Hood, *Adv. Immunol.* **44**, 1 (1989).
6. T. C. Südhof, J. L. Goldstein, M. S. Brown, D. W. Russell, *Science* **228**, 815 (1985); T. C. Südhof *et al.*, *ibid.*, p. 893.
7. R. A. F. Dixon *et al.*, *Nature* **321**, 75 (1986); T. Kubo *et al.*, *ibid.* **323**, 411 (1986).
8. M. Marchionni and W. Gilbert, *Cell* **46**, 133 (1986); W. Gilbert, M. Marchionni, G. McKnight, *ibid.*, p. 151.
9. K. Obaru, T. Tsuzuki, C. Setoyama, K. Shimada, *J. Mol. Biol.* **200**, 13 (1988); C. Setoyama, T. Joh, T. Tsuzuki, K. Shimada, *ibid.* **202**, 355 (1988); M. C. Shih, P. Heinrich, H. M. Goodman, *Science* **242**, 1164 (1988); F. Quigley, W. F. Martin, R. Cerff, *Proc. Natl. Acad. Sci. U.S.A.* **85**, 2672 (1988).
10. If the exons have a general probability distribution P_i (for the ith exon), then the expectation of doubles is $n(n-1)/2$ times ΣP_i^2 since ΣP_i^2 is the total probability that a pair matches. Similarly triples are $[n(n-1)(n-2)/6]\ \Sigma P_i^3$. If the distribution were exponential $P(x=(1/\sigma)\exp(-x/\sigma)$, then the estimate for the universe is 2σ, the number of exons that would account for 86 percent of the occurrences.
11. C. Burks *et al.* "GenBank: Current Status and Future Direction," in *Molecular Evolution: Computer Analysis of Protein and Nucleic Acid Sequences*, R. F. Doolittle, Ed. (Academic Press, New York, in press); in our work, exons were drawn from the GenBank (version 56) and EMBL (version 15) databases.
12. (RATMABPA) K. Drickamer and V. McCreary, *J. Biol. Chem.* **262**, 2582 (1987); M. E. Taylor, P. M. Brickell, R. K. Craig, J. A. Summerfield, *Biochem. J.* **262**, 763 (1989).
13. (HUMFIXG) D. M. Anson *et al.*, *EMBO J.* **3**(5), 1053 (1984); S. Yoshitake, B. G. Schach, D. C. Foster, E. W. Davie, K. Kurachi, *Biochemistry* **24**, 3736 (1985).
14. (HUMCFXII) D. E. Cool and R. T. A. MacGillivray, *J. Biol. Chem.* **262**, 13662 (1987).
15. (HUMC1QB1) K. B. M. Reid, *Biochem J.* **231**, 729 (1985).
16. (MMIIGC) N. Koch, W. Lauer, J. Habicht, B. Dobberstein, *EMBO J.* **6**, 1677 (1987).
17. (RABCN) M. E. Fini, I. M. Plucinska, A. S. Mayer, R. H. Gross, C. E. Brinckerhoff, *Biochemistry* **26**, 6156 (1987).
18. (MUSMUPBS) A. J. Clark, P. M. Clissold, R. Al Shawi, P. Beattie, J. Bishop, *EMBO J.* **3**, 1045 (1984); A. J. Clark, P. Ghazal, R. W. Bingham, D. Barrett, J. O. Bishop, *ibid.* **4**, 3159 (1985).
19. (CRECPSBA) J. M. Erickson, M. Rahire, J.-D. Rochaix, *EMBO J.* **3**, 2753 (1984); J. K. Mohana Rao, P. A. Hargrave, P. Argos, *FEBS Lett.* **156**(1), 165 (1983).
20. (MUSBAND3I) R. R. Kopito, M. A. Andersson, H. F. Lodish, *Proc. Natl. Acad. Sci. U.S.A.* **84**, 7149 (1987).
21. (HUMTNFB) G. E. Nedwin, *Nucleic Acids Res.* **13**, 6361 (1985).
22. (RATRHL) J. O. Leung, E. C. Holland, K. Drickamer, *J. Biol. Chem.* **260**, 12523 (1985).
23. The full-protein alignments were made with the algorithm of Needleman and Wunsch, as implemented in the GAP programs provided by the University of Wisconsin [J. Devereux, P. Haeberli, O. Smithies, *Nucleic Acids. Res.* **12**(1), 387 (1984)]. Where necessary, alignments were anchored on the exon pair we identified.
24. G. H. Swift *et al.*, *J. Biol. Chem.* **259**, 14271 (1984).
25. P. J. O'Hara *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **84**, 5158 (1987).
26. S. K. Hanks, A. M. Quinn, T. Hunter, *Science* **241**, 42 (1988)
27. C. C. Hyde, S. A. Ahmed, E. A. Padlan, E. W. Miles, D. R. Davies, *J. Biol. Chem.* **263**, 17857 (1988).
28. Another way of doing this calculation is to consider the excess in each percentage match and calculate a chi-squared ((observed-expected)^2/expected) for each of the entries (pooling values first for the small entries) then add all of these chi-square

values and demand significance for the number of degrees of freedom corresponding to the number of entries pooled. Both of these calculations suggest that the excess is significant at well above the 95 percent level.
29. T. A. Jones and S. Thirup, *EMBO J.* **5**, 819 (1986).
30. R. Unger, D. Harel, S. Wherland, J. L. Sussman, *Proteins: Struct. Funct. Genet.* **5**, 355 (1989).
31. C. Sander and R. Schneider, *Proteins*, in press.
32. (HUMCG1A1) K. S. E. Cheah, N. G. Stoker, J. R. Griffin, F. G. Grosveld, E. Solomon, *Proc. Natl. Acad. Sci. U.S.A.* **82**, 2555 (1985).
33. (HUMAPOB1) T. J. Knott *et al.*, *Science* **230**, 37 (1985).
34. (HUMEGFRG) S. Ishii *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **82**, 4920 (1985).
35. (HUMC1PA) M.-L. Chu *et al.*, *Nature* **310**, 337 (1984).
36. (HUMEL) Z. Indik *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **84**, 5680 (1987).
37. (GGHSP108) M. Forsgren, B. Raden, M. Israelsson, K. Larsson, L.-O. Heden, *FEBS Lett.* **213**, 254 (1987).
38. (HUMNFLG) J.-P, Julien *et al.*, *Biochim. Biophys. Acta* **909**, 10 (1987).
39. (SCO1G2) J. J. M. Dons *et al.*, *EMBO J.* **3**, 2102 (1984).
40. (HUMFN) A. R. Kornblihtt, K. Vibe-Pedersen, F. E. Baralle, *Nucleic Acids Res.* **12**, 5853 (1984).
41. (GGCFPSE) C.-C. Huang, C. Hammond, J. M. Bishop, *J. Mol. Biol.* **181**, 175 (1985).
42. (MUSCOLA2) M. Kurkinen, M. P. Bernard, D. P. Barlow, L. T. Chow, *Nature* **317**, 177 (1985).
43. (BTTHYR) J. Parma, D. Christophe, V. Pohl, G. Vassart, *J. Mol. Biol.* **196**, 769 (1987).
44. (BMOCH11A) K. Iatrou, S. G. Tsitilou, F. C. Kafatos, *Proc. Natl. Acad. Sci. U.S.A.* **81**, 4452 (1984).
45. (MUSKETEPI) P. M. Steinert, R. H. Rice, D. R. Roop, B. L. Trus, A. C. Steven, *Nature* **302**, 794 (1983); T. M. Krieg *et al.*, *J. Biol Chem.* **260**, 5867 (1985).
46. (HUMALBGC) P. P. Minghetti *et al.*, *J. Biol. Chem.* **261**, 6747 (1986).
47. (HUMKEREP) D. Marchuk, S. McCrohon, E. Fuchs, *Proc. Natl. Acad. Sci. U.S.A.* **82**, 1609 (1985).
48. We thank J. Willett, H. Spencer, and R. C. Lewontin for helpful statistical advice, J. Knowles for careful reading of the manuscript, and members of the Gilbert Lab for useful assistance and criticism. This work is supported by NIH grant GM37997-03.

24 August 1990; accepted 22 October 1990



" Calm down, Helen. We've been the focus of watch-dog groups before."