ELSEVIER

Gene 205 (1997) 141–144

GENE
AN INTERNATIONAL JOURNAL ON GENES AND GENOMES

The correlation between introns and the three-dimensional structure of proteins[1]

Sandro Jose de Souza, Manyuan Long, Lloyd Schoenbach, Scott William Roy, Walter Gilbert *

The Biological Laboratories, 16 Divinity Avenue, Cambridge, MA 02138, USA

Accepted 5 June 1997

Abstract

We test the hypothesis that introns were used to construct the first genes from small exons, whose protein products represent compact elements of structure. For any three-dimensional structure, a computer program analyzes the structure into a set of modules, segments of the polypeptide chain bounded in space by a maximum diameter, separated by a set of 'boundary regions'. The 'boundary regions' are such that if the gene were divided by an intron in each 'boundary region', the protein would be divided into modules less than the specified diameter. Using a set of 32 ancient proteins, which have no introns in prokaryotes, we examine the intron positions in their eukaryotic homologs and show that the introns are correlated with modules of diameter 21, 28 and 33 Å, with P values below 0.001. © 1997 Elsevier Science B.V.

Keywords: Exons; Module; Evolution

1. Introduction

The two contrasting views, that introns arose early, to be used to assemble the first genes (Doolittle, 1978; de Souza et al., 1996a), or that introns appeared late in evolution (Palmer and Logsdon, 1991; Stoltzfus et al., 1994), junk DNA breaking up previously continuous genes, embody radically different views of the origin of genes. An introns-late view is often justified by a phylogenetic argument, by observing that there are no introns in the prokaryotes, that there are no introns in many simple unicellular eukaryotes, and that a full flourish of introns appears in the higher plants and the vertebrates. An introns-early view could be supported by arguing that the splicing machinery is an RNA-enzyme based, looking like an original function, and patterns of intron loss can be easily seen. Recently novel arguments with great statistical force have been developed that support the early introns hypothesis (Long et al., 1995; de Souza et al., 1996b).

The argument that we will focus on here is the conjecture that introns existed from the beginning of evolution where their role was to enhance recombination between exons in order to facilitate the first creation of genes. One form of this hypothesis, 'The Exon Theory of Genes' (Gilbert, 1987), stated explicitly that the first genes were essentially small open reading frames, 15 or 20-amino-acid-long products from small segments of genetic material. These small polypeptide chains served first as aggregates to be the first protein enzymes and were then assembled into genes made up of small exons. The basic process of evolution on this picture is to shuffle exons to make novel genes (Gilbert, 1978). The evolutionary processes are then recombination within introns, the sliding and drift of introns to change the peptide sequence around the splicing site, and the loss of introns to make more complex exons, to be shuffled in turn. One easy way of losing introns to make more complex exons is a known process, retroposition, in which a gene with an intron/exon structure produces a spliced RNA copy. Then, a reverse transcription copies that RNA back into DNA, and the DNA reinserts into the chromosome. If that reinsertion is into the intron of a previously-existing gene, that DNA copy can now become a complex exon in a new gene (a complete example of this process is the *Jingwei* gene in *Drosophila* (Long and Langley, 1993)).

This hypothesis suggests that complex exons have appeared over time; hence, the search for original exons

* Corresponding author. Tel.: +1 617 4950760; Fax: +1 617 4964313; e-mail: gilbert@chromo.harvard.edu

[1] Presented at the International Society of Molecular Evolution Meeting, Guanacaste, Costa Rica, 6–10 January 1997.

PII S0378-1119(97)00401-0

is difficult. Since the distribution of current exons is peaked near lengths of 40 amino acids (Long et al., 1995), there would have been two to three fusions of exons, on the average, to make today's pattern.

The Exon Theory of Genes predicts that proteins will turn out to be made up of simple elements, compact units of folding or function, called 'modules' by Mitiko Go (1981), which in turn will be eventually shown to be related to the exons. This paper shows that there is strong statistical support for this prediction.

By 'module', we mean a segment of a polypeptide chain that folds back upon itself in space in such a way that all of the C-alpha positions can be circumscribed by some fixed diameter. Roughly speaking, that means that that segment lies inside a sphere of that specified diameter. The most common diameter used historically is 28 Å.

If exons correspond to compact elements, then the introns should define the boundaries of such modules. Over the years, this notion was first used by Mitiko Go to predict the existence of an intron in globin (Go, 1981) and then was used to predict the existence of introns in triosephosphate isomerase (Straus and Gilbert, 1985; Gilbert et al., 1986). However, the difficulty with this concept has been that the 'boundary' of a module is not well defined. Since, in general, the 'spheres' that define modules overlap, there is no unique position for a boundary (or an intron) defined along the polypeptide backbone. We have turned this problem into a virtue (de Souza et al., 1996b). If we cannot make a precise prediction for the boundary, we can make a clear prediction of a 'boundary region.' We take the entire region that lies in each overlap between the rough spheres that define modules as possible sites for introns and so define this overlap region as a 'boundary region.' Thus, we define 'boundary regions' that have the property that if one were to put an intron into each of those regions, the protein structure would be dissected into modules of the specified size. Immediately, one has now specified the problem in a way that yields a simple statistical test. These 'boundary regions' correspond to about a third of the protein sequence, and the test, obviously, is to ask if there is an 'excess' of intron positions in these 'boundary regions.' Clearly, the alternative model, in which introns have been added to the DNA by some process over the course of evolution, predicts that the introns would be added randomly to the DNA, and thus, of course, not favor these 'boundary regions.'

We can give these 'boundary regions' a computationally simple definition. Fig. 1 shows a plot, a plot first introduced by Mitiko Go (Go, 1981), which displays the distances between all pairs of amino acids in a protein. By marking all pair distances that are more than 28 Å by black, we can see along the diagonal triangular areas which are the delineating black regions.

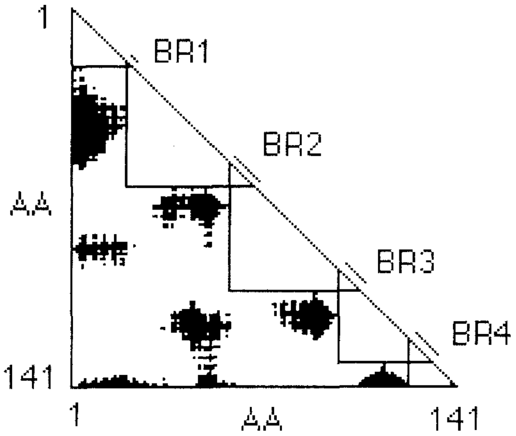

Fig. 1. Black regions in this Go plot represent pairs of alpha-carbons that are separated by 28 Å or more in horse hemoglobin. The triangles along the diagonal define the modules. The 'boundary regions' are defined as the region of overlap of those triangles. If at least one intron were to be placed in each 'boundary region', the protein would be dissected into five modules each less than 28 Å in diameter.

These are the modules at 28 Å. The five large triangles correspond to the largest possible modules at 28 Å, and the overlaps between these triangles define the 'boundary regions'. A computer program, INTER-MODULE, does essentially this calculation for any protein.

To construct a statistical test for the correspondence of intron positions and modules, we selected a set of ancient proteins of known three-dimensional structures. These 'ancient' proteins have bacterial counterparts, which have no introns, and have homologs in the eukaryotes with introns. We examined a set of 32 such proteins and we found, originally, in the database a corresponding set of 570 intron positions. (We count each different intron position along the nucleotide sequence of these genes (de Souza et al., 1996b).)

Now, on any introns-late model, all of these positions have to represent introns added to the previously-existing genes. On all those models, there could be no exon shuffling in the history of these specific genes, since they are continuous in bacteria, and all such models believe that such genes came into existence in the progenote as continuous structures. Thus, for all introns-late models, these introns are derived properties. Furthermore, these intron positions could not be related to any exon shuffling or to any alternative splicing, in these models, since the prokaryotic and eukaryotic homologs are conserved and colinear. However, the introns-early model conjectures that some of these introns may have come from the original genes and that the bacterial genes have lost the introns.

Using the dataset of 570 intron positions, we expect about 182 to lie in the 'boundary regions,' but we find 214. That is a 17% excess. While not a gigantic excess, since there are so many positions, we find $\chi^2 = 8.2$ and $P = 0.005$. Our first question is answered in a dramatic fashion. For a standard-sized module, picked for historical reasons at 28 Å, there is a clear, significant excess of introns in the 'boundary regions' (de Souza et al., 1996b).

Could some type of biased addition model save the argument for the introns-late view? Is there something special about these 'boundary regions'? We have examined these sequences to ask if there is any bias that might serve as a reason for special intron targeting. Is there any DNA sequence bias? So far none that we can see. Is there any amino acid sequence bias? Not that we have detected. Occasionally sequences such as AGG or AGGT are suggested to be target sequences, presplicing sequences, for intron addition. There is no bias for these sequences, or even the sequence GG, in the 'boundary regions' compared to the rest of the protein. Although there has been the conjecture that introns lie preferentially on the surface of proteins (Craik et al., 1982), we did not see any surface bias in our dataset (de Souza et al., 1996b). Introns lie on the surface with the same frequency that amino acids in the protein lie on the surface.

One last piece of evidence is even more suggestive. Consider a subset of putative 'ancient' introns, introns that have common positions across great evolutionary distances, that have common positions in three out of the four of plants, fungi, invertebrates and vertebrates. There are 20 introns in this category. Thirteen lie in the 'boundary regions', whereas only 6.5 are expected. This is a 100% excess, a much higher excess than for the full set. Is this 100% excess in the set of 20 statistically different from the 17% excess in the set of 570? Yes, the χ^2 for this comparison is 6.48 with a $P=0.01$. This group, picked as a group of putative ancient introns, is more biased toward the 'boundary regions' than the general set. That is exactly what one expects if the general set of intron positions contains both ancient introns as well as a background of moved or added introns.

Now, this analysis can be taken further, because the

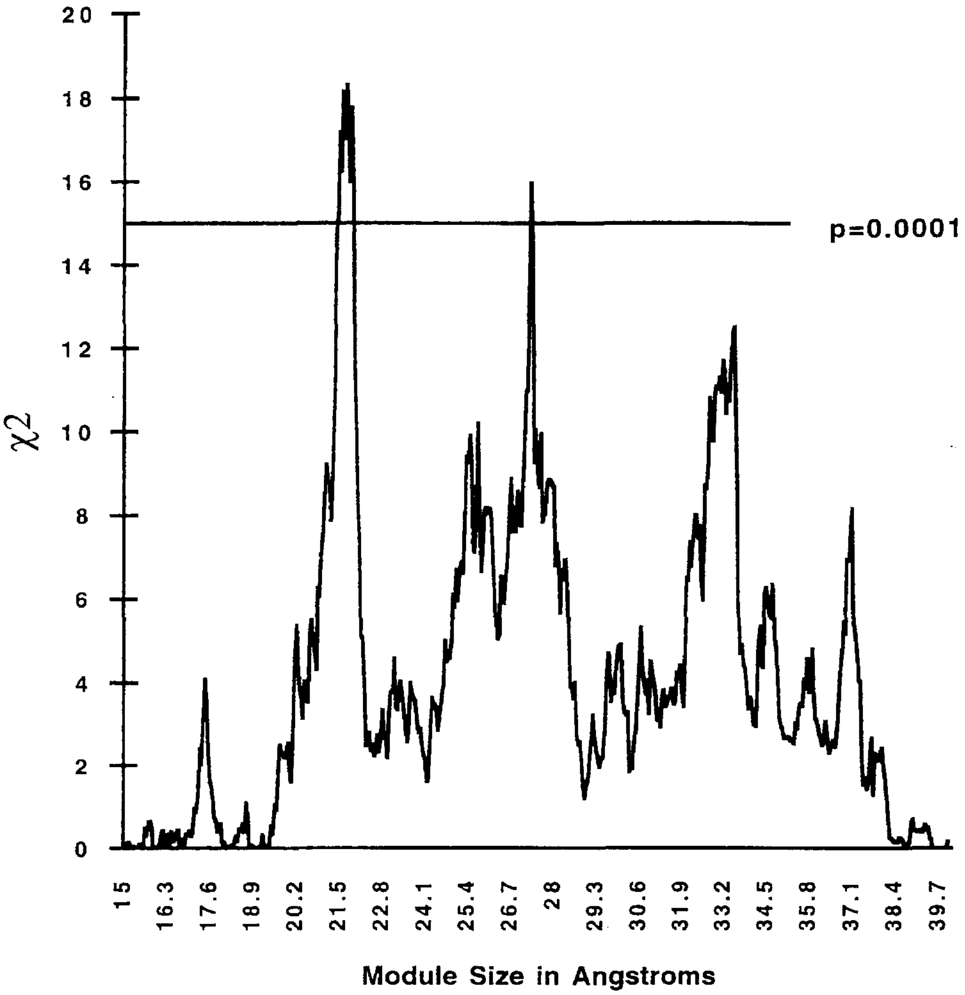

Fig. 2. The χ^2 values for the distribution of intron positions in relation to the 'boundary regions' defined from 15 to 40 Å in intervals of 0.05 Å. There are three major significant peaks of excess of introns in the 'boundary regions' around 21 Å ($P<0.0001$), 27 Å ($P<0.0001$) and 33 Å ($P<0.0005$).

computer program will define such 'boundary regions' for any diameter. When we ran the program with arbitrary diameters and plotted the output as a χ^2 for the fit of introns to the 'boundary' regions, we got three peaks of significance, each in the $P=0.001$ range (de Souza et al., 1996b). These three peaks of statistical significance correspond to diameters near 21 Å, 28 Å and 33 Å. For these sizes, there is a correlation between intron positions and module boundaries. We interpret this result as meaning that some introns are correlated with 21 Å modules, some with 28 Å modules and some with 33 Å. Superficially, one might worry about the statistics since, in effect, we have carried out almost 1000 calculations along the way, and one might expect one of those to show a fluctuation 1000-fold away from the mean. However, the phenomenon is robust, not random fluctuations, since one can see these peaks in the actual excess of intron positions as a function of module diameter.

Fig. 2 shows the current state of this calculation. By using a later version of the database based on GenBank, version 96, there are now 662 intron positions rather than the 570 used before. The three peaks are accentuated by the greater number of introns. The peak at 21 Å has now reached a $\chi^2=19$. The additional ancient intron positions in the database mostly come from the *C. elegans* project, which appears to be enriched in ancient introns.

What is the meaning of these peaks? The computer program effectively divides up the proteins by defining regions for hypothetical introns. Another way of looking at this is to say that the program has divided up the protein into a set of hypothetical exons. What are the sizes of these exons? In these terms, this pattern makes a little more sense. The three peaks, 21 Å, 28 Å and 33 Å, correspond roughly to average lengths of modules of 15 amino acids, 22 amino acids and 30 amino acids, respectively. This correlation makes two statements: first, that the ancient introns did define elements of structure of the protein and, secondly, that these original elements were of the size hypothesized by The Exon Theory of Genes, about 15, 22 and 30 residues long.

The future direction of this analysis is to try to understand the actual pattern of modules. The Exon Theory of Genes predicts that the relevant modules predicted by INTER-MODULE will have been used by exon shuffling, over and over again, to create these proteins. Thus, we expect to be able to identify a set of shapes among the 21, 28 and 33 Å modules that are reused and, in toto, account for a major part or all of protein structure. We expect that as we do this analysis, we will also see that the introns correlate most relevantly with these reusable modules. Lastly, we plan to study the amino acid sequences of such sets of 'reused' modules to see if any trace of conservation can be possibly detected.

This last work is in its infancy. At this point, all that we know is that these modules often contain turns (to make them compact). The most common repeated form at 21 Å is an alpha helix followed by a turn and a strand.

2. Conclusions

We have presented a strong argument that at least some of the introns that we see today are very ancient. One consequence of that argument is to suggest that the first proteins were assembled by introns used to connect small exons.

Acknowledgement

We thank the National Institutes of Health, grant no. GM 37997, for support. S.J. de Souza was supported by Fundacao de Amparo a Pesquisa do Estado de Sao Paulo (FAPESP-Brasil) and the PEW-Latin American Fellows Program.

References

Craik, C.S., Sprang, S., Fletterick, R., Rutter, W.J., 1982. Intron–exon splice junctions map at protein surfaces. Nature 299, 180–182.

de Souza, S.J., Long, M., Gilbert, W., 1996a. Introns and gene evolution. Genes to Cells 1, 493–505.

de Souza, S.J., Long, M., Schoenbach, L., Roy, S.W., Gilbert, W., 1996b. Intron positions correlate with module boundaries in ancient proteins. Proc. Natl. Acad. Sci. USA 93, 14632–14636.

Doolittle, W.F., 1978. Genes in pieces: Were they ever together? Nature 272, 581–582.

Gilbert, W., 1978. Why genes in pieces? Nature 271, 501

Gilbert, W., Marchionni, M., McKnight, G., 1986. On the antiquity of introns. Cell 46, 151–154.

Gilbert, W., 1987. The exon theory of genes. Cold Spring Harbor Symp. Quant. Biol. 52, 901–905.

Go, M., 1981. Correlation of DNA exonic regions with protein structural units in haemoglobin. Nature 291, 90–93.

Long, M., Langley, C., 1993. Natural selection and the origin of *jingwei*, a chimeric processed functional gene in *Drosophila*. Science 260, 91–95.

Long, M., Rosenberg, C., Gilbert, W., 1995. Intron phase correlations and the evolution of the intron/exon structure of genes. Proc. Natl. Acad. Sci. USA 92, 12495–12499.

Palmer, J.D., Logsdon, J.M.J., 1991. The recent origins of introns. Curr. Opin. Genet. Dev. 1, 470–477.

Stoltzfus, A., Spencer, D.F., Zuker, M., Logsdon, J.M.J., Doolittle, W.F., 1994. Testing the exon theory of genes: The evidence from protein structure. Science 265, 202–207.

Straus, D., Gilbert, W., 1985. Genetic engineering in the Precambrian: Structure of the chicken triosephosphate isomerase gene. Mol. Cell. Biol. 5, 3497–3506.