Intron distribution difference for 276 ancient and 131 modern genes suggests the existence of ancient introns

Alexei Fedorov*, Xiaohong Cao*†, Serge Saxonov*‡, Sandro J. de Souza§, Scott W. Roy*, and Walter Gilbert*¶

*Department of Molecular and Cellular Biology, Harvard University, 16 Divinity Avenue, Cambridge, MA 02138; and §Laboratory of Computational Biology, Ludwig Institute for Cancer Research, São Paulo Branch, Rua Professor Antonio Prudente, 109-4. andar, CEP 01509-010, São Paulo, Brazil

Contributed by Walter Gilbert, September 18, 2001

Do introns delineate elements of protein tertiary structure? This issue is crucial to the debate about the role and origin of introns. We present an analysis of the full set of proteins with known three-dimensional structures that have homologs with intron positions recorded in GenBank. A computer program was generated that maps on a reference sequence the positions of all introns in homologous genes. We have applied this program to a set of 665 nonredundant protein sequences with defined three-dimensional structures in the Protein Data Bank (PDB), which yielded 8,217 introns in 407 proteins. For the subset of proteins corresponding to ancient conserved regions (ACR), we find that there is a correlation of phase-zero introns with the boundary regions of modules and no correlation for the phase-one and phase-two positions. However, for a subset of proteins without prokaryotic counterparts (131 non-ACR proteins), a set of presumably modern proteins (or proteins that have diverged extremely far from any ancestral form), we do not find any correlation of phase-zero intron positions with three-dimensional structure. Furthermore, we find an anticorrelation of phase-one intron positions with module boundaries: they actually have a preference for the interior of modules. This finding is explicable as a preference for phase-one introns to lie in glycines, between G | G sequences, the preference for glycines being anticorrelated with the three-dimensional modules. We interpret this anticorrelation as a sign that a number of phase-one introns, and hence many modern introns, have been inserted into G | G ''protosplice'' sequences.

There are two opposing points of view about the origin of introns and their role in the evolution of early genes. The ''introns-early'' theory suggests that introns arose at the earliest steps of evolution to create the first genes by the exon shuffling of small pieces (1, 2). Conversely, an ''introns-late'' point of view suggests that the earliest genes were created by processes that led to continuous long genes, and that exon shuffling occurs only late in evolution in the genes for the complex eukaryotes. One of the main arguments to support an introns-early theory has been the observation of a correlation between intron positions and structural features of the encoded proteins. In 1981, Mitiko Go suggested (3) that primitive exons might correspond to compact modules in the three-dimensional protein structure. A similar argument for the correlation of introns with module boundaries was used for triosephosphate isomerase (4, 5). Other reports have suggested the correspondence of introns to structural features of proteins (2, 6–10). Nonetheless, a simple one-to-one correspondence of exons to protein units has many exceptions: many introns are found within protein domains or other protein structural units.

The question of the origin of ''ancient'' genes is most easily studied by considering ancient conserved proteins, i.e., proteins that are homologous between prokaryotes and eukaryotes. The genes for such proteins have no introns in the prokaryotes but have introns in the complex eukaryotes. The introns-late viewpoint suggests that the continuous genes in the prokaryotic homologs correspond to the ancestral forms, and that the introns in the eukaryotic forms have all been added. Conversely, the introns-early viewpoint suggests that some or all of these introns might be ancient; they have been involved in exon shuffling in the history of these gene regions because they are collinear between the prokaryotes and the eukaryotes. Stoltzfus *et al*. (11) found no correspondence of intron positions to the structural protein elements for four ancient genes. Logsdon (12) studied the intron distribution in triosephosphate isomerase and argued that the position of introns did not support a correlation with exon structure, although a different conclusion was arrived at by Gilbert and Glynias (13). Other workers also have argued for introns-late (14, 15).

These earlier investigations dealt with only a few proteins, and the argument often turned on whether statistical significance had been reached. The extensive worldwide sequencing taking place as part of the genome project has made it possible now to test large groups of proteins with very extensive surveys of intron positions. A group of 44 ancient proteins was analyzed by de Souza *et al.* (9, 10) and by Roy *et al.* (16) using two different definitions of modules, which are compact regions of protein structure. These analyses found that only introns that lie in phase zero, between the codons, show a correlation with module boundaries. Introns that lie after the first or second base of a codon, in phase one or phase two, do not show any correlation. These papers concluded that about a third of the phase-zero intron positions in these ancient genes might represent introns that were involved in the exon shuffling process that was used to assemble these genes. de Souza *et al.* (9) defined modules as compact regions of polypeptide chains whose α-carbon distances were all less than a defined diameter and which lie within a sphere of that diameter, and they examined the correlation of introns with the overlap between such spheres. Roy *et al.* (16) used a centripetal definition of modules; i.e., the minima in a function that averages the sum of the squares of the distances between α-carbons to define the boundaries of modules. Both groups show a strong and extremely significant correlation of the phase-zero intron positions with module boundaries. Furthermore, Roy *et al.* (16) showed that the subset of those phase-zero introns that are ancient by phylogenetic arguments, being similar in position between different phylogenetic kingdoms, are more closely correlated with the module boundaries than are the rest of the phase-zero intron positions.

To go further, this paper presents an analysis of the full set of proteins with known three-dimensional structures that have

Abbreviations: EID, exon/intron database; ACR, ancient conserved regions.

†Present address: Genzyme Corporation, 5 Mountain Road, Framingham, MA 01701.

‡Present address: Stanford Medical Informatics, 251 Campus Drive, Medical School Office Building X-215, Stanford, CA 94305.

¶To whom reprint requests should be addressed. E-mail: gilbert@nucleus.harvard.edu.

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

homologs with intron positions recorded in GenBank. Because we must examine hundreds of proteins and thousands of intron positions, this analysis can be done only in a completely computerized fashion. We wrote a computer program, INTRONMAP, which will map on a reference sequence the positions of all introns in homologous genes. We have applied this program to a set of 665 nonredundant protein sequences with defined three-dimensional structures in the Protein Data Bank (PDB). For the subset of proteins corresponding to ancient conserved regions (ACR), we find that there is a correlation of phase-zero introns with the boundary regions of modules and no correlation for the phase-one and phase-two positions. However, for a subset of proteins without prokaryotic counterparts, a set of presumably modern or very altered proteins, we do not find any correlation of phase-zero intron positions with three-dimensional structure. Furthermore, for this subset of proteins, we find an anticorrelation of phase-one intron positions with module boundaries: they actually have a preference for the interior of modules. This finding is explicable as a preference for phase-one introns to lie in glycines, between G | G sequences, the preference for glycines being anticorrelated with the three-dimensional modules. We interpret this result as a sign that a number of phase-one introns, and hence many modern introns, have been inserted into G | G protosplice sequences.

Materials and Methods

The Largest Protein Sample with Known Three-Dimensional Structure. Hobohm and Sander (17) defined a nonredundant (25% homology) list of proteins with known three-dimensional structure from the PDB (obtained from the web site, ftp://ftp.embl-heidelberg.de/pub/databases/protein_extras/pdb_select). We selected from this list a protein sample of 665 sequences that met two criteria: (*i*) sequences were longer than 100 amino acids, and (*ii*) there were no gaps in the α-carbon coordinate data longer than three amino acids.

Intron Mapping on a Set of Proteins. An exon/intron database (EID) (18), derived from the GenBank release 116, was the source of introns. We used the pEID file of the EID database containing the protein sequences, as well as positions and phases of all introns. The set of 665 protein sequences was aligned with all protein sequences from the pEID file by using the gapped BLAST 2.0 program (19). Based on these alignments, intron positions from the EID entries were mapped onto the reference sequences with the help of the INTRONMAP program, written in Perl and executed on a Dec-α station 500 running Digital UNIX 4.0b. Conditions for intron mapping were the following: (*i*) the threshold for the alignment of E-values between the reference protein and EID protein entries was 10^{-4}; (*ii*) each intron could be mapped only once; and (*iii*) introns with gaps in the alignment in the vicinity of their positions were not included (see web page, http://mcb.harvard.edu/gilbert/INTRONMAP).

From the initial 665 protein set, introns have been mapped successfully on 407 proteins (for the alignment threshold of E-value $\leq 10^{-4}$). These 407 sequences were divided first into two groups: 237 ancient proteins with at least one eukaryotic-prokaryotic homologous pair with a BLAST 2.0 E-value less than 10^{-4}, and a remaining group of 170. Running Position-Specific Iterated-BLAST on the 170-protein group further identified another 39 proteins with more distant prokaryotic homologues. These 39 proteins were combined with the 237 ancient proteins as the ACR group of 276 proteins. The remaining 131 proteins are the non-ACR group.

Sample with Artificial Intron Positions. For the simulation experiment, we prepared samples of artificial intron positions located between the guanines of GG or AGG sites of the reference coding sequences of the 276 ACR proteins (G | G sample or AG | G sample, respectively).

Correlation Between Intron Positions and Protein Module Structure. The INTERMODULE program of de Souza *et al.* (10) was used for this analysis.

Results

Analysis of Intron Positions on a Previous Set of 44 Ancient Proteins. We adjusted the parameters of the program INTRONMAP, which maps intron positions on a reference-protein sequence, to reproduce as closely as possible the previous results on a set of 44 ancient proteins by de Souza *et al.* (10). Fig. 1*A* shows the correlation between 1,061 intron positions, obtained by INTRONMAP, and the module boundary regions, where Fig. 1*B* shows the previous data from de Souza *et al.* (10). In the new set of 1,061 introns and the previous set of 1,073 from de Souza *et al.* (10), 15% of the intron positions differ because of the difference in the intron selection criteria between the two mapping approaches. Nonetheless, the set of intron positions obtained by INTRONMAP has the same pattern of correlation with module boundaries as the previous set.

Correlation of Intron Positions and Protein Module Structure for the Full Set of Proteins with Known Three-Dimensional Structure. We applied INTRONMAP to the set of 665 nonredundant proteins with known three-dimensional structure; it successfully mapped introns on 407 proteins from the initial list of 665. We then divided these 407 proteins into two groups: 276 ACR proteins, for which there exists at least one pair of eukaryotic-prokaryotic homologous protein pairs, and a residual 131 non-ACR group. Fig. 2 shows the correlation of the intron positions with the boundary regions of modules for the ACR and the non-ACR groups. The distributions of intron positions for the 44 ancient proteins (Fig. 1) and the 276 ACR proteins (Fig. 2*A*) have similar properties. Both groups have an excess of phase-zero introns in the boundary regions for the whole range of module sizes. There are some differences between these two sets of proteins as well. The excess of phase-zero introns for the 276 ACR group is statistically significant ($\chi^2 = 13.1$, $P = 3.0 \times 10^{-4}$) for the module size of 27–28 Å. Although statistically significant peaks for intron excess in the boundary regions exist for modules with size 21 Å ($\chi^2 = 12.6$, $P = 3.9 \times 10^{-4}$) and 32 Å ($\chi^2 = 12.1$, $P = 5.0 \times 10^{-4}$) for the group of 44 ancient proteins, they are not so prominent for the 276 ACR group. For both groups of ancient proteins, the phase-one intron positions appear in the boundary regions less than the random expectation.

Fig. 2 shows a remarkable difference in the distribution of introns between the 276 ACR proteins and the 131 non-ACR proteins. There is no correlation between phase-zero intron positions and module boundaries for the 131 non-ACR proteins. Furthermore, the avoidance of phase-one introns for the module boundaries is more prominent for the non-ACR proteins for the whole spectrum of module sizes. The avoidance peaks around modules 12–25 Å in diameter, and it is statistically very significant (max χ^2 is 20.9, $P = 4.8 \times 10^{-6}$).

There are only weak correlations of phase-two intron positions and module boundaries, which change from positive to negative at different module sizes and can be explained as random fluctuations.

The Influence of the Parameters of the Intron-Mapping Program on Intron Distribution. We investigated the effect of choosing different criteria for the selection of introns in the INTRONMAP program on the correlation between the resulting intron positions and the Go-module (3) boundaries. Variation in the "alignment threshold" in the range from 10^{-4} to 10^{-15} for the BLAST E-values and variations in the threshold for the "local

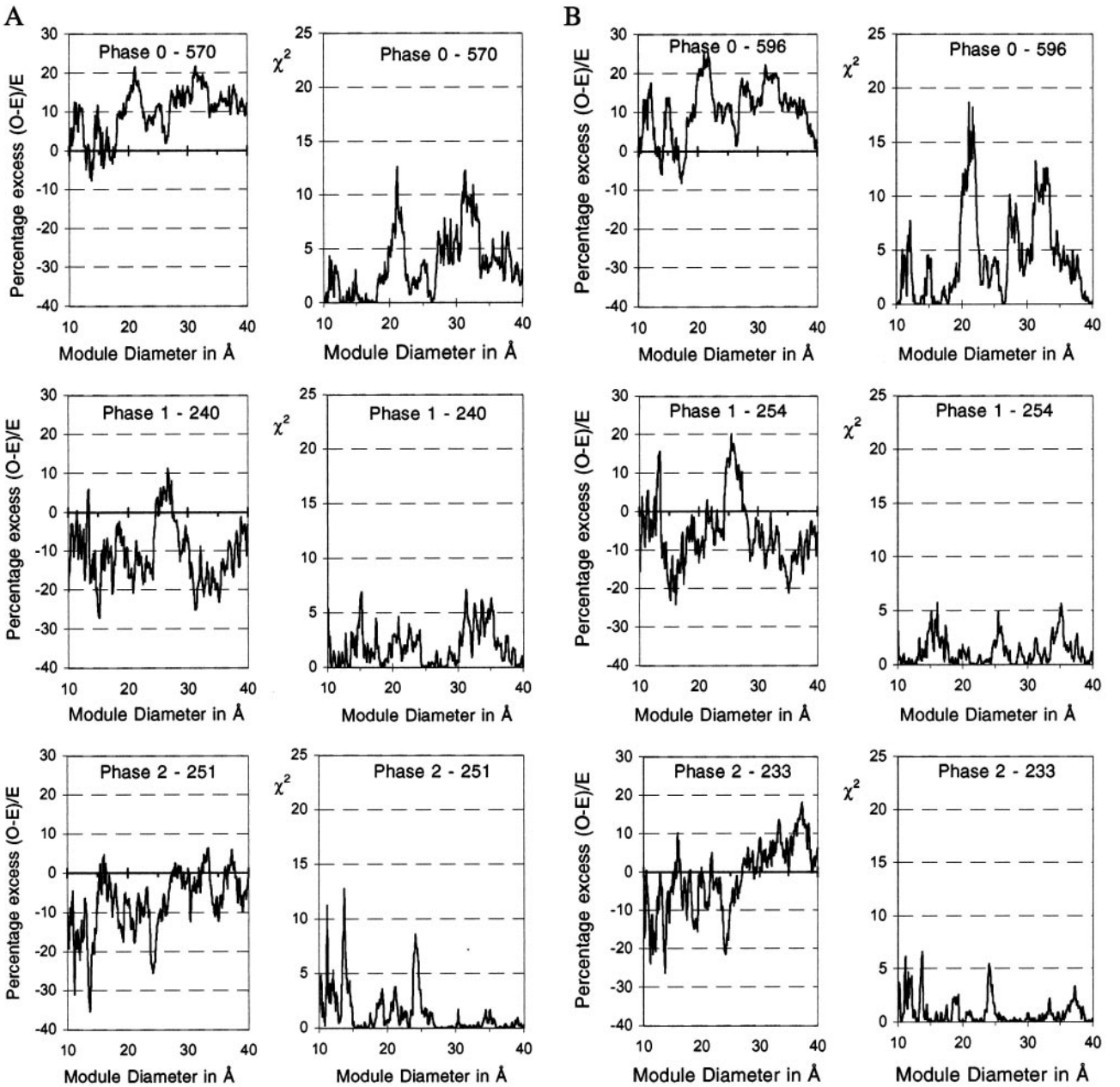

Fig. 1. Correlation of intron positions with module boundaries of 44 ACR proteins. The percentage of excess/depletion of intron positions above expectation [(O − E)/E × 100%] and χ^2 values for the excess [degrees of freedom (DF) = 1]. (*A*) Current data set of intron positions obtained by INTRONMAP. Number of analyzed introns: phase zero, 570; phase one, 240; phase two, 251. (*B*) Previous data set of intron positions (10). Number of analyzed introns: phase zero, 596; phase one, 254; phase two, 233.

homology" in the vicinity of the introns in the range from 5 to 30% caused only minor effects on the correlations of intron with the module boundaries.

Changing the use of the "gap" parameter made the greatest difference. The removal of those introns that have "local gaps" in the alignments in the vicinity of their positions significantly increases the correlation of phase-zero introns with module boundaries. The presence of such gaps in alignments is the main source of errors in intron mapping and can lead to rather arbitrary placements. Without such removal of introns with local gaps, the intron sample would have a bias: an excess of intron positions in low-conservation regions.

Species Effects on Intron Distributions. de Souza *et al.* (10) showed the exclusion of all vertebrate intron positions from the group of 44 ancient genes increased the correlation of phase-zero introns with the module boundaries. However, in the larger 276-member group of ACR proteins and the set of 3,328 phase-zero intron positions, eliminating the vertebrate intron positions to leave 2,851 phase-zero positions did not affect the correlation (data not shown). Presently, a considerable portion of introns in GenBank come from the genomic sequencing projects and were predicted by a computer program without experimental supporting evidence. One of the largest group of nonexperimentally determined introns in GenBank belong to *C. elegans*. Some of the computer predicted *C. elegans* intron/exon structures are wrong (20). The exclusion of the introns from *C. elegans* from our data set led to an increase in the correlation of phase-zero intron positions with module boundaries: the χ^2 increased by a factor

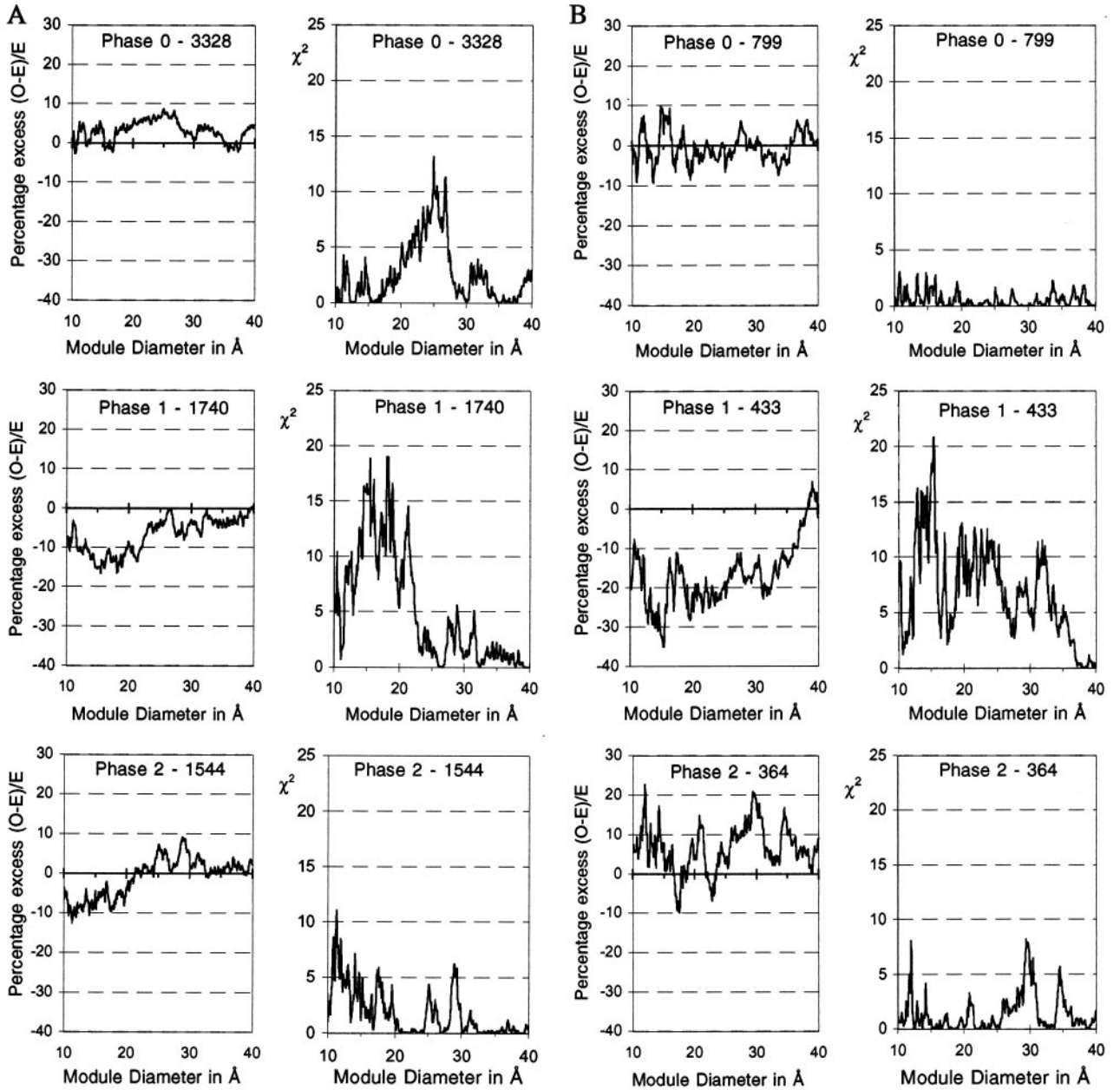

Fig. 2. Correlation of intron positions with module boundaries of 276 ACR and 131 non-ACR proteins. The percentage of excess/depletion of intron positions above expectation [(O − E)/E × 100%] and χ^2 values for the excess (DF = 1), which was calculated for the data set of intron positions obtained by INTRONMAP. (*A*) 276 ACR proteins. Number of analyzed introns: phase zero, 3,328; phase one, 1,740; phase two, 1,544. (*B*) 131 non-ACR proteins. Number of analyzed introns: phase zero, 799; phase one, 433; phase two, 364.

of two (max χ^2 is 22.6, $P = 2.0 \times 10^{-6}$ for modules 25 Å in diameter, Fig. 3). The removal of the *C. elegans* introns caused no effect on the distribution of phase-zero introns in the 131 non-ACR group (results not shown). The presence of *C. elegans* introns has no significant effect on the distribution of phase-one and phase-two introns in both ACR and non-ACR groups (results not shown). Because many of the *C. elegans* introns are computer-predicted and may be in error, we analyzed these introns further. *C. elegans* intron positions were divided into two groups, computer-predicted vs. experimentally confirmed; both groups had similar patterns for the phase-zero distributions (data not shown).

Why Do the Phase-One Introns Avoid the Module Boundary?. There is a considerable deficit of phase-one intron positions in the boundary regions of both ACR and non-ACR proteins. One possible explanation of this phenomenon is that 41% of phase-one introns lie in Gly residues, and Gly, being the smallest amino acid residue, more often lies in the interior of compact modules. To test this hypothesis, we examined what the distribution of introns would be if they lay between G | G pairs, or if they lay in AG | G sequences. We took the coding sequences for the reference sequence of each of the 276 ACR proteins and examined the distribution of G | G patterns with respect to module boundaries. We calculated the preference for the boundary region (and the χ^2 for all of the 17,266 G | G pairs. Fig. 4 shows that if introns were to lie in G | G sequences, there would be no particular preference pattern for phase zero or phase two. However, for phase one, we see the strong avoidance of module

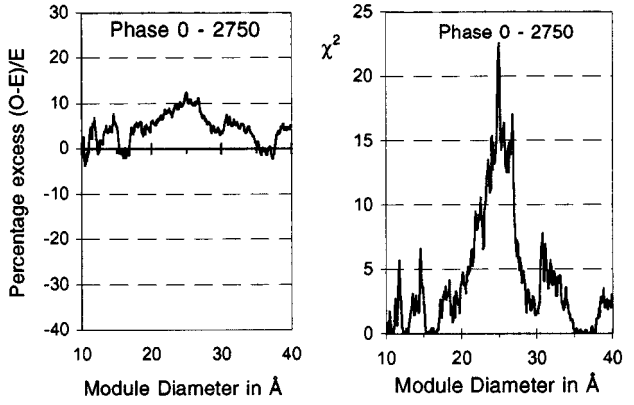

Fig. 3. Correlation of intron positions (without *C. elegans* introns) with module boundaries of 276 ACR proteins. The percentage of excess/depletion of intron positions above expectation [(O − E)/E × 100%] and χ^2 values for the excess (DF = 1), which was calculated for the data set of intron positions obtained by INTRONMAP. Number of analyzed phase-zero introns, 2,750.

boundaries. The distribution of artificial G | G introns in boundaries of different size modules has the same pattern and is even more profound than that of the real phase-one introns. Computer simulations showed that when 40–50% of phase-one artificial introns are within G | G sites and the rest of them are distributed randomly, the distribution of phase-one artificial and real introns in module boundaries have the same patterns (data not shown).

Discussion

Our comprehensive computer analysis of the correlation of intron positions with Go-module structures revealed a complex picture of intron/exon evolution. On the largest set of 276 ancient nonredundant proteins, we confirmed a previous observation, made on a group of 44 ancient proteins (10, 16), that phase-zero introns are preferably located within module boundaries. The study of the previously unexplored group of nonancient genes, specific for eukaryotes, gave new results. The phase-zero introns of 131 non-ACR genes do not correlate with protein-module structure. These facts support the assumption that there is a subgroup of ancient phase-zero introns. These ancient introns preferably locate inside module boundaries and, presumably, were involved in the exon-shuffling process. The non-ACR proteins are most likely proteins whose sequences and gene structures are so modified by evolution that any similarity to ancestral forms was lost.

Phase-one introns avoid module boundaries in the large sample of 276 ACR genes, and this avoidance is even stronger for the 131 non-ACR genes (about 25% deviation with $\chi^2 \approx 15$–20, $P \approx 1.0 \times 10^{-4} - 7.7 \times 10^{-6}$; Fig. 2*B*). This result was unexpected. However, phase-one introns frequently occur within Gly codons, and because Gly is the smallest amino acid and of particular importance for protein folding, there is a structural basis for the suggestion that Gly codons lie inside Go-modules.

Why should phase-one introns prefer to lie specifically in Gly codons, GGN? Possibly, introns were inserted between two guanidines, into the so-called protosplice sites of genes, at late stages of evolution. The recent study of splicing sites of different organisms by Long *et al.* (21, 22) showed that the best candidates for protosplice sites are G | G or AG | G sequences. If the hypothesis of intron insertions inside G | G or AG | G sites is true, then introns inserted within phase one must occur only inside Gly codons GGN. We tested the consequences of intron insertion into protosplice sites in simulation experiments. A

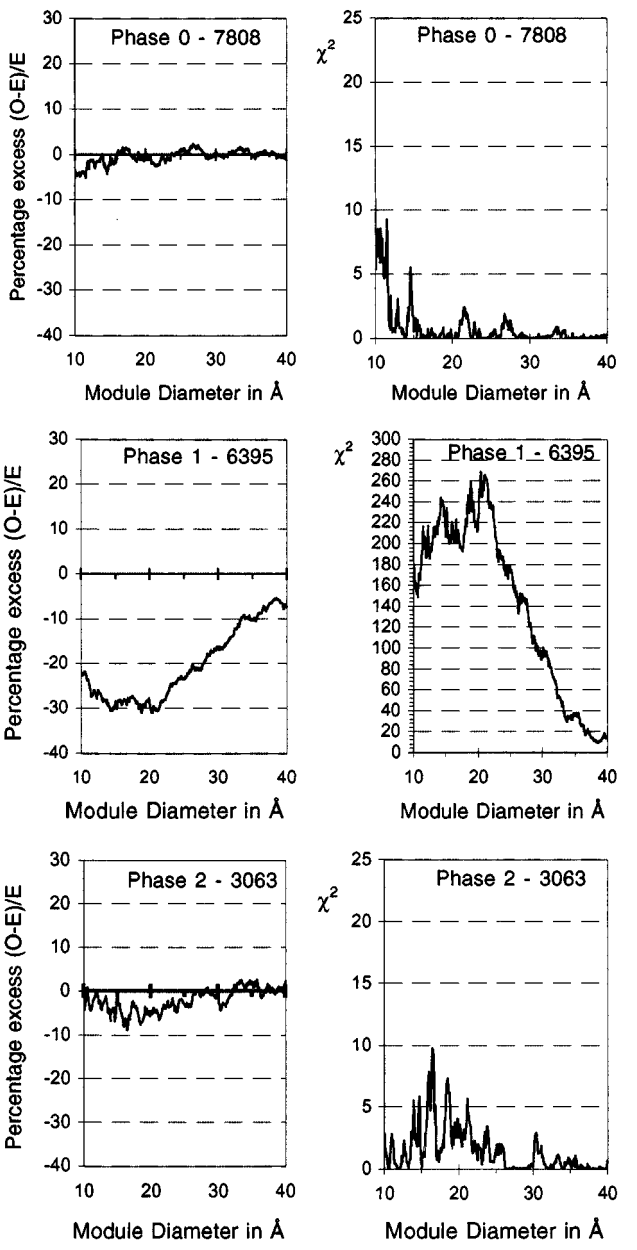

Fig. 4. Correlation of artificial (G | G) intron positions with module boundaries of 276 ACR proteins. The percentage of excess/depletion of intron positions above expectation [(O − E)/E × 100%] and χ^2 values for the excess (DF = 1). Number of analyzed artificial introns: phase zero, 7,808; phase one, 6,395; phase two, 3,063.

computer program generated samples of artificial introns located only inside G | G or AG | G sites of the reference sequence of this set of genes. Then, the correlation of artificial introns with module boundaries was analyzed in the same way as a sample of real introns by using the program INTERMODULE. The artificial phase-zero introns show no correlation with module boundaries. However, the artificial phase-one introns, all in Glys, are sharply anti-correlated (about a 30% effect). Fig. 4 shows that the avoidance of module boundaries is similar for the phase-one artificial introns and phase-one real introns (Fig. 2). Thus, we conclude that the avoidance of phase-one introns of module boundaries is caused by the preferential location of phase-one introns inside Gly codons.

These data on the comparison of artificial and real intron distributions shed a new light on intron evolution. One important result is that phase-zero artificial introns of the 276 ACR genes, like the real phase-zero introns of the 131 non-ACR genes, do not correlate with module boundaries. This result shows that the preference of phase-zero real introns for the module boundaries in the ancient genes is not caused by intron insertions. It is more likely that a group of ancient introns exists in the ACR genes. We presume that these ancient introns have strong correlation to module boundaries as a result of their involvement in the process of exon shuffling.

The avoidance of phase-one introns for the module boundaries is less prominent in the ACR genes than in their non-ACR counterparts. This difference could be explained through postulating the existence of a small fraction of ancient phase-one introns in ACR genes: a preference of ancient introns for module boundaries would compensate for the strong effect of boundary avoidance from the newly inserted introns. Endo *et al.* (unpublished data) studied the distribution of intron positions inside codons for evolutionary invariable amino acids, which are identical among all available prokaryotic and eukaryotic homologous sequences from GenBank. These authors showed that 70% of evolutionarily invariable codons interrupted by phase-one introns code for Gly. The remaining 30% of the invariable codons interrupted by phase-one introns code for another 11 amino acids. These phase-one introns interrupting non-Gly invariable codons could not appear recently by the insertions inside G | G sites. Thus, perhaps about 30% of phase-one introns inside ACR genes might be ancient.

We conclude that the intron–exon structure had an intricate history, and that different stages of exon and intron evolution have occurred. We found strong evidence for the existence of ancient introns. Presumably, this group of ancient introns shows a preference to lie inside the protein module boundaries because of their previous involvement in exon shuffling. At the same time, we view the majority of present-day introns as ''modern'' introns, inserted recently into the G | G protosplice sites. The negative correlation of modern phase-one introns with module boundaries is due to their presence in protosplice sites within Gly codons. The existence of both ancient introns and modernly inserted introns can explain their observed distributions in ACR and non-ACR gene samples.

We thank Dr. Manyuan Long and Dr. W. Ford Doolittle for valuable critiques of the manuscript.

1. Doolittle, W. F. (1978) *Nature (London)* **272,** 581–582.
2. Gilbert, W. (1987) *Cold Spring Harbor Symp. Quant. Biol.* **52,** 901–905.
3. Go, M. (1981) *Nature (London)* **291,** 90–93.
4. Marchionni, M. & Gilbert, W. (1986) *Cell* **46,** 133–141.
5. Tittiger, C., Whyard, S. & Walker, V. K. (1993) *Nature (London)* **361,** 470–472.
6. Stone, E. M., Rothblum, K. N. & Schwartz, R. J. (1985) *Nature (London)* **313,** 498–500.
7. Michlelson, A. M., Blake, C. C. F., Evans, S. T. & Orkin, S. H. (1985) *Proc. Natl. Acad. Sci.* USA **82,** 6965–6969.
8. Duester, G., Jornvall, H. & Hatfield, G. W. (1986) *Nucleic Acids Res.* **14,** 1931–1941.
9. de Souza, S. J., Long, M., Schoenbach, L., Roy, S. W. & Gilbert, W. (1996) *Proc. Natl. Acad. Sci.* USA **93,** 14632–14636.
10. de Souza, S. J., Long, M., Klein, R. J., Roy, S., Lin, S. & Gilbert, W. (1998) *Proc. Natl. Acad. Sci.* USA **95,** 5094–5099.
11. Stoltzfus, A., Spencer, D. F., Zuker, M., Logsdon, J. M. J. & Doolittle, W. F. (1994) *Science* **265,** 202–207.
12. Logsdon, J. M., Tyshenko, M. G., Dixon, C., Jafari, J. D., Walker, V. K. & Palmer, J. D. (1995) *Proc. Natl. Acad. Sci.* USA **92,** 8507–8511.
13. Gilbert, W. & Glynias, M. (1993) *Gene* **135,** 137–144.
14. Rzhetsky, A., Ayala, F. J., Hsu, L. C., Chang, C. & Yoshida, A. (1997) *Proc. Natl. Acad. Sci.* USA **94,** 6820–6825.
15. Cho, G. & Doolittle, R. F. (1997) *J. Mol. Evol.* **44,** 573–584.
16. Roy, S. W., Nosaka, M., de Souza, S. J. & Gilbert, W. (1999) *Gene* **238,** 85–91.
17. Hobohm, U. & Sander, C. (1994) *Protein Sci.* **3,** 522–524.
18. Saxonov, S., Daizadeh, I., Fedorov, A. & Gilbert, W. (2000) *Nucleic Acids Res.* **28,** 185–190.
19. Altschul, S. F. Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D. J. (1997) *Nucleic Acids Res.* **25,** 3389–3402.
20. Goto, O. (1998) *Mol. Biol. Evol.* **15,** 1447–1459.
21. Long, M., de Souza, S. J., Rosenberg, C. & Gilbert, W. (1998) *Proc. Natl. Acad. Sci.* USA **95,** 219–223.
22. Long, M. & Rosenberg, C. (2000) *Mol. Biol. Evol.* **17,** 1789–1796.