Proc. Natl. Acad. Sci. USA
Vol. 88, pp. 2380–2383, March 1991
Biochemistry

Splice junctions follow a 205-base ladder

(intron/exon/chromatin/higher order structure)

J. S. BECKMANN* AND E. N. TRIFONOV†

*Department of Plant Genetics and Breeding, Institute of Field and Garden Crops, Agricultural Research Organization, The Volcani Center, P.O. Box 6, Bet Dagan 50250, Israel; and †Department of Polymer Research, The Weizmann Institute of Science, Rehovot 76100, Israel

Communicated by Donald M. Crothers, December 19, 1990 (received for review March 26, 1990)

ABSTRACT **Of all aspects of mRNA maturation the accuracy of intervening sequence excision and exon ligation is, perhaps, the most enigmatic. Attempts to identify the essential elements involved in this process have thus far not yielded any satisfactory answer as to what structural (sequence) features are prerequisite for the vital precision of this process. In our search for underlying structural orders we asked whether exons and introns had any positional preferences within a gene. This analysis led to the unexpected discovery that the DNA length is synchronized between successive 3′ splicing sites as well as between successive 5′ splicing sites, with a frame of ≈205 base pairs. This observation reveals additional organization of genes in eukaryotes and, perhaps, links gene splicing with chromatin structure.**

Several properties differentiate eukaryote from prokaryote cell; notable among these differences are chromatin structure and gene splicing. Trying to find a rationale governing gene splicing in eukaryotes, we realized that chromatin structure and gene splicing might somehow be related—for example, the necessity for better protection of exons and exon–intron boundaries might predicate their involvement in the nucleosomes. On the other hand, giving priority to chromatin and, particularly, to its higher-order structure, some adjustments in gene sequences would be expected—e.g., introduction of introns to allow formation of appropriate higher-order structure (1). Nucleosome positions are known to be largely DNA sequence-dependent (2); their relative orientations in space appear dictated by sequence as well (3, 4). The sequences of uninterrupted structural genes with their coding capacity might not always be compatible with the complex set of rules underlying specific nucleosome positioning (5). We believe sequence requirements of the higher-order chromatin structure to be equally important in this respect. Thus introns, by interrupting the coding sequences at certain positions, could render the protein-coding and chromatin sequence requirements compatible. Once introduced, the intervening sequences might then serve other purposes as well. We, therefore, examined intragenic structural relationships of exon–intron boundaries to check whether these boundaries follow any positional preferences, perhaps reflecting a splicing–chromatin connection. In this communication we report that spliced genes do obey a regular spatial organization; such a connection is thus indicated.

MATERIALS AND METHODS

The collection of human and murine sequences examined was retrieved from the European Molecular Biology Laboratory data base (release 15). The following intragenic distances were recorded: (*i*) from every 5′ junction site to its downstream 3′ junction site (intron length) and to the nearest and all subsequent downstream 5′ junction sites; (*ii*) from every 3′ junction site to the next 5′ junction site (exon length) and to the nearest and all subsequent downstream 3′ junction sites. Entries were carefully examined to avoid duplications. All genes of simple reiterated motifs, such as collagen, elastase, as well as pseudogenes, were also discarded. The resulting human data base consisted of 155 genes and a total of 530 exons and 518 introns. Splice sites and their positions were used as listed in the feature tables of the sequences. Only a small proportion of splice sites in the data base are authentic—i.e., confirmed by mRNA or protein sequences, whereas others (consensus sites) are accepted as they were estimated by the corresponding authors. The data were analyzed in bulk (grouping authentic splice sites and questionable consensus splice sites together). Frequency distributions were obtained by counting the sequences in groups of increased lengths; successive groups differed from one another by 20 bases. Histograms were generated by "smoothing" (see below) the data.

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.

RESULTS AND DISCUSSION

Histograms indicate the measured distances separating like 5′–5′ (Fig. 1*A*) and 3′–3′ (Fig. 1*B*) ends of the human sequences. A rather regular profile is seen in both cases, which survives after summation of the two histograms (Fig. 1*C*), thus indicating a common periodical component. Notwithstanding the noise, maxima appear periodically, approximately every 200–210 base pairs (bp) as far as 2000 bp downstream. The first two peaks are most pronounced. The decreased amplitude as a function of increased length is but a trivial consequence of the fact that more short genes than long genes occur in the data base. Initially these results were obtained on a much smaller data set than that used for Fig. 1. Because of the unexpected character of the observed regularity, we examined a larger collection of sequences, which confirmed our first observation. Moreover, the same results were obtained when samples of the data were examined and counted by hand (to exclude any flaws in the computer program) and when the analysis involved batching the sequences at increments of 10 bases instead of 20 bases and smoothing every 10 bases, within windows of 60 and 90 bases, rather than 100 bases (data not shown).

A murine data set of comparable size was assembled and analyzed similarly; the same precautions were used as described for human sequences. Examination of the murine data set produced the results shown in Fig. 2 *A–C*, where, once again, regular underlying periodical patterns could be seen for 5′–5′ and 3′–3′ distances, with peaks occurring every 200–210 bases. The underlying periodicity, common for the human and mouse sequences, is seen most clearly when all histograms (human 5′–5′, 3′–3′ and mouse 5′–5′, 3′–3′) are summed (Fig. 3). Fourier analysis of these profiles yields periods of 205, 208, and 206 bases, indistinguishable within error bar (± 2), for the human, murine, and total data sets, respectively. The calculated amplitude of this oscillation is

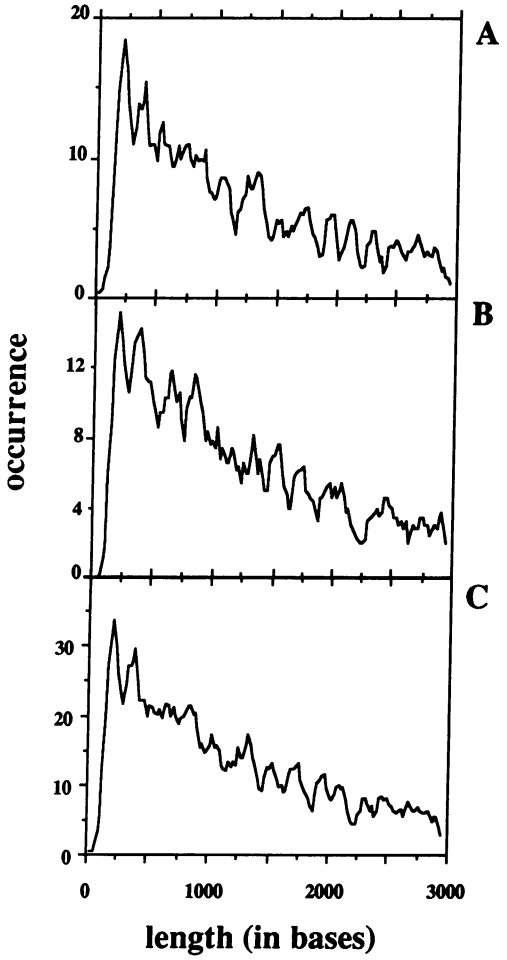

FIG. 1. Histograms of the distances separating like ends 5′–5′ (*A*) and 3′–3′ (*B*) and the sum of these two (*C*) from human gene sequences retrieved from European Molecular Biology Library; (release 15; see below). Distances within a gene were recorded from every position to its next downstream like position and subsequent like positions. Distances up to 3000 bases were used in these analyses; they were grouped in size clusters with increments of 20 and counted. The histograms were derived upon smoothing in 20-base steps with a window of 100 bases—thus, the figures along the ordinate correspond to a smoothed average of actual occurrences of the distances within 20 bases. Human entries used in analysis (EMBL release 15) include the following: HS242G2, HSA1ATP, HSACCYBB, HSACHR2, HSACTGA, HSACTH, HSADAG, HSAFL12, HSAFL34, HSAFPCP, HSAGL1, HSALBGC, HSALDC, HSALP5, HSAMYA2, HSANFG1, HSANTCD8, HSAPC3A, HSAPOA2G, HSAPOA4A, HSAPOAIA, HSAPOC2G, HSAPRT, HSB2M2, HSBGL3, HSBGPG, HSBLYM1, HSBNGF, HSBSF2, HSCATF, HSCATG7, HSCD14G, HSCD2G1, HSCEANCA, HSCFOS, HSCFVII, HSCFXII3, HSCG01, HSCG1A10, HSCRELA, HSCRPG, HSCRYGX1, HSCS1, HSCSFGMA, HSCYP450, HSDHFR01, HSERBBR, HSERPA, HSEV15VK, HSFABP, HSFBRGG, HSFERG2, HSFESFPS, HSFIBEDA, HSFIXG, HSFN3A, HSFUR1, HSFVII, HSGAS1, HSGASTA, HSGCSFG, HSGLUCG2, HSGMCSFG, HSGOA1, HSGSHPXG, HSGSTPIG, HSHCDCB, HSHER2B, HSHL07, HSHLAB27, HSHLADPB, HSHLADZA, HSHLASBA, HSHLIA, HSHP201, HSHPARS1, HSHSC70, HSHSP27, HSHST, HSIAIG2, HSIAIG4, HSIG05, HSIGF2G, HSIGGC3, HSIGHAM, HSIGHBC, HSIGHTC2, HSIGK15, HSIGLPAV, HSIGV126, HSIL05, HSIL1AG, HSIL1B, HSIL2B, HSINSPR, HSINT1G, HSKER672, HSKER673, HSKIN10, HSLACTG, HSLCATG, HSLMYC1, HSMDR1G, HSMED, HSMETIE, HSMHC1, HSMHCGE2, HSMHCP42, HSMHDC3B, HSMHSXA, HSMIS, HSMPO1, HSMT1B1, HSMYCC, HSNFLG, HSNMYC, HSOPS, HSOTNPI, HSPCRF, HSPLAPL, HSPLPSPC, HSPPPA, HSPRCA, HSPRPH1, HSREN03, HSREN04, HSREP10, HSRPBG1, HSRPS14, HSSAA, HSSB3B46, HSSBA2P, HSSLIPG, HSSOMI, HSTCBCB, HSTCBV85, HSTCBYT, HSTCRAC, HSTCRT3D, HSTGFBG4, HSTHB, HSTHY1A, HSTHYR8, HSTKRA, HSTNFAB, HSTNFB, HSTPA, HSTUBAG, HSTUBB2, HSUK, HSUPA, HSVK02, HSVPNP, HSYG01, and HSYUBGI.

≈4 SDs over background, which can also be estimated from the actual occurrences presented in Fig. 3.

Both intron and exon lengths are known to vary extensively. The distributions of both show single maxima around 110 bp tailing up to as much as 1500 bp for exons and to several thousand bp for introns (data not shown; see also ref. 6). Nothing in these single-peak distributions would indicate the effect presented in Figs. 1–3. Even when single exon–intron pairs are taken (i.e., the nearest 5′–5′ and 3′–3′ pairs only), a second peak is observed (Fig. 4), which must reflect the basic periodic organization of the interrupted genes.

Remarkably, both 5′ and 3′ junctions, apparently independently from one another, follow the same ladder. Several peaks are seen in both cases at multiples of ≈205 bases, except for the peak at position three, which fails to appear in the total sum. Recall that all splice sites (both authentic and consensus) listed in the European Molecular Biology Laboratory data bank were used in this analysis. The fact that not all claimed sites might be *bona fide* splice sites would only blur the image and, hence, reduce the effect. Thus, the observed periodicity indicates that the actual effect is probably even stronger than reported here.

Apart from the periodicity, perhaps even more surprising is the persistence of the pattern over long distances, some 2000 bases away. These results seem to imply that splicing junctions, or some structural context thereof, have a memory of the presence and position of other neighboring junctions, being preferable in $n \times 205$-base phase with them. Yet, we emphasize that positioning of splice sites on an almost equidistant basis is far from the general rule for all junctions. Many short exons are followed by either a very short or a very long intron (and *vice versa*), thus not conforming to the pattern described. Thus, these facts did not lead to the *a priori* anticipation of such regular organization. In this context the preferential 205-base frame comes as a surprise.

These observations definitely deserve explanation, and several hypotheses come to mind. The most attractive hypothesis is suggested by the fact that the 205-base ladders are reminiscent of and quantitatively close to "nucleosomal repeats" (7). Projecting the observation on other eukaryotes, in general, is difficult because there is insufficient data to analyze genes of other organisms. Those organisms with atypical chromatin (yeast) or splicing organization (yeast, *Drosophila*) might lack the regularity seen in human and mouse genes. Should exon–intron junctions reflect a nucleosomal organization? One possible connection is the protective role nucleosomes might assume while packing the splice junctions. Splice sites differ from many other structural and regulatory DNA sequence elements by their stringent exactitude; misfiring of a single base in one or another of the canonical GU and AG ends of the introns would be disastrous for the entire molecule of mature mRNA. Splice sites are therefore extremely vulnerable to mutations. The great care through which these sequences are protected from mutations can also be seen in the elegant study of ref. 8. It thus seems likely that means to protect these sites from mutational hazards might have been devised throughout evolution, in particular, by encompassing these sites within a more protective environment, such as found in nucleosomes. This strategy, of course, would result in preferential positioning of the splice sites at the nucleosome repeat distance—i.e., avoiding the lengths that would put one of the junctions in a linker region.

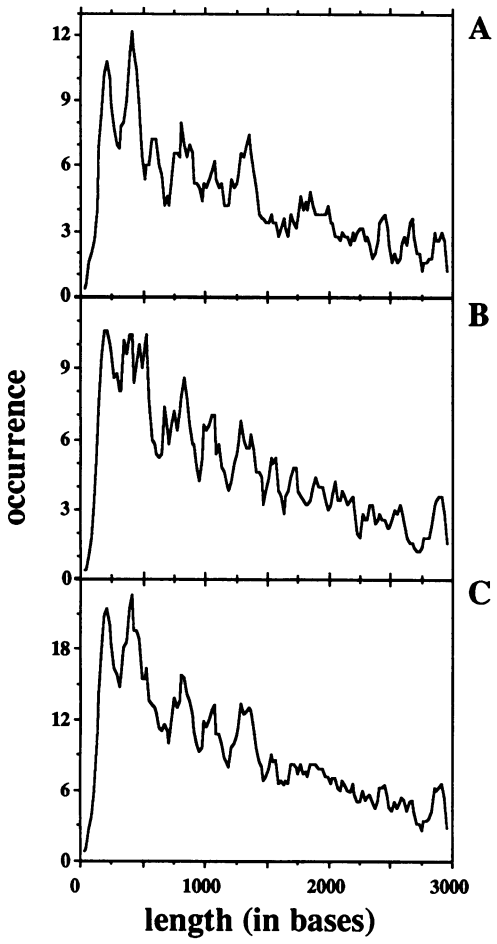

FIG. 2. Histograms of the distances separating like ends 5′–5′ (*A*) and 3′–3′ (*B*) and the sum of these two (*C*) from murine gene sequences retrieved from European Molecular Biology Library (release 15). Distances within a gene were recorded and analyzed as explained for Fig. 1. Mouse entries used in analysis (EMBL release 15) include the following: MMABLC1A, MMACASA, MMADIG, MMALDH1, MMAMY2, MMANF, MMANT12, MMAPOIVA, MMAPRT, MMASP, MMASPAT6, MMBAND3I, MMCAIIN, MMCD42, MMCD43, MMCD46, MMCFOS, MMCRYG2D, MMCRYS, MMCY02, MMCY03, MMELA2G, MMERPA, MMG37, MMG4CRY1, MMGCSFG, MMGFAPD, MMGFBPA3, MMGK5, MMGMCSFG, MMGPD, MMGSHPX, MMH19G, MMH2KK, MMH2LD, MMHBAZ, MMHOX13, MMIEBB3, MMIG13, MMIG15, MMIG36, MMIGEHC, MMIGEHC1, MMIGG1, MMIGG10G, MMIGVH9, MMIIGB, HSMDR1G, MMIIGC, MMIL01, MMIL1BG, MMIL3G, MMIL4G12, MMIL5G, MMINSIIG, MMINT2, MMKALL, MMLDHAG, MMLYT212, MMLYT22, MMMETII, MMMH04, MMMHC4A1, MMMHCAB1, MMMHCQ1, MMMHCQ3, MMMHIEAD, MMMHTLA1, MMMHTLA2, MMMLC134, MMMUPBS6, MMMYCE12, MMMYCNA, MMMYOGG2, MMNFIL, MMNFMG, MMNGFA2, MMNMYC, MMNMYC5, MMNUCLEO, MMODC2, MMPIM1, MMPROT, MMPRPMPB, MMRKAL3, MMRPL30, MMRPL3A, MMRPOII2, MMRPS16, MMSAA1B, MMTCA2, MMTCBJB2, MMTCBVB, MMTHY1GC, MMTLAC, and MMTNFAB.

In fact, the extreme relative vulnerability of splice junctions leads one to wonder why apparently useless introns were not lost during evolution. The contrary seems to have occurred. We are then forced to conclude that introns are far from ancillary and probably occupy some essential functions. We believe that this function may also be related to a structural requirement for chromatin folding. Both nucleosome positions along DNA and their folding in space appear largely dictated by the nucleotide sequences (2, 4). Consequently, the arrangement of nucleosomes, organized by the genomic sequence with alternating exons and introns, has to be substantially different from that expected were noninterrupted protein-coding DNA sequence to organize the chromatin structure. Being spliced, the mRNA sequence carries neither the originally encoded chromatin design nor any positionally and spatially consistent chromatin design at all, serving primarily in its protein-coding capacity. Thus, gene splicing can be viewed as a device for simultaneously satisfying conflicting protein-coding and chromatin-organization sequence requirements, as originally suggested in earlier studies (1, 5). Nucleosomes and their higher-order organization could constitute one of the synchronizing factors for the observed preferentially regular spacing of splice sites along a gene, as well as the *raison d'être* of introns.

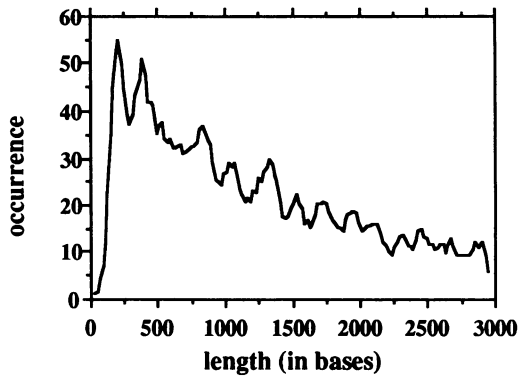

FIG. 3. Cumulative integrated histogram of the distances separating all human and murine intron ends (5′–5′ and 3′–3′ together).

Other phasing mechanisms could also be involved. Units of ≈200 bp of DNA are seen in other instances: Okazaki pieces are synthesized in segments of about this size before being end-ligated (9). mRNA processing could also structurally involve a kind of processive scanning "measure" that would accommodate ≈200 nucleotides at a time. The packaging of the mRNA precursors in the heterogeneous nuclear RNP particles might also follow some structural regularity (10) to position the splice junctions properly for efficient and accurate processing. And, of course, the observed regularity could result from joint pressures of all the above size preferences.

Whatever mechanism is considered, it also must account for the fact that rather long-distance peaks are seen as well. We assume the presence of some responsible phasing elements in the DNA sequences, but preliminary analyses have failed to recognize any simple nucleotide signal; perhaps more complex periodically organized patterns await detection.

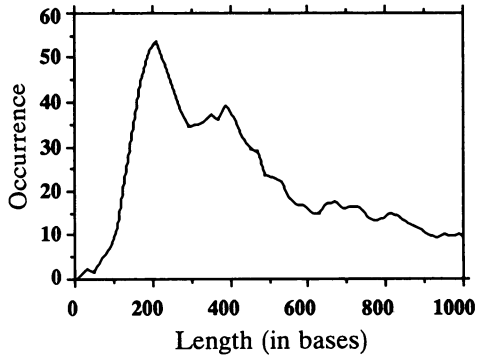

FIG. 4. Histogram of the distances separating single exon-intron pairs, nearest like ends (5′–5′ and 3′–3′) from the human and murine data sets. Distances within each gene were recorded and analyzed as explained for Fig. 1.

The expert help in data assembly by Dr. Philip Bucher is highly appreciated. This work is contribution no. 2840.E (1989 series) from the Agricultural Research Organization, The Volcani Center, Bet Dagan, Israel.

1. Zuckerkandl, E. (1981) *Mol. Biol. Rep.* **7**, 149–158.
2. Mengeritsky, G. & Trifonov, E. N. (1983) *Nucleic Acids Res.* **11**, 3833–3851.
3. Noll, M., Zimmer, S., Engel, A. & Dubochet, J. (1980) *Nucleic Acids Res.* **8**, 21–42.
4. Ulanovsky, L. E. & Trifonov, E. N. (1986) in *Biomolecular Stereodynamics III*, eds. Sarma, R. H. & Sarma, M. H. (Adenine, Guilderland, NY), pp. 35–44.
5. Soloviev, V. V. & Kolchanov, N. A. (1986) *Dokl. Biochem. (Engl. Transl.)* **284**, 286–290.
6. Hawkins, J. D. (1988) *Nucleic Acids Res.* **16**, 9893–9908.
7. van Holde, K. E. (1988) *Chromatin* (Springer, New York).
8. Smith, C. W. J., Porro, E. B., Patton, J. G. & Nadal-Ginard, B. (1989) *Nature (London)* **242**, 243–247.
9. Ogawa, T. & Okazaki, T. (1980) *Annu. Rev. Biochem.* **49**, 421–457.
10. Barnett, S. F., Friedman, D. L. & LeStourgeon, W. M. (1989) *Mol. Cell. Biol.* **9**, 492–498.