Structure of the 32-kDa Galectin Gene of the Nematode Caenorhabditis elegans*

(Received for publication, April 17, 1997, and in revised form, July 14, 1997)

Yoichiro Arata‡, Jun Hirabayashi, and Ken-ichi Kasai
From the Department of Biological Chemistry, Faculty of Pharmaceutical Sciences, Teikyo University, Sagamiko, Kanagawa 199-01, Japan

Galactins are a family of soluble β-galactoside-binding lectins distributed in both vertebrates and invertebrates and, more recently, found also in fungi. The 32-kDa galectin isolated from the nematode Caenorhabditis elegans (Hirabayashi, J., Satoh, M., and Kasai, K. (1992) J. Biol. Chem. 267, 15485–15490) was the first “tandem repeat-type” galectin, containing two homologous carbohydrate-binding sites. Here, we report the structure of the nematode 32-kDa galectin gene. Physical mapping by yeast artificial chromosome polytene filter hybridization revealed that the 32-kDa galectin gene is located on chromosome II. Analysis of the transcript (1.4 kilobases) showed the presence at its 5'-end of a 22-nucleotide spliced leader sequence (SL1). The entire genomic structure spanning >5 kilobase pairs (kbp), including the 5'-noncoding region, two intervening sequences (introns 1 and 2), and the 3'-noncoding region, was completely determined by the combination of genomic polymerase chain reaction and conventional colony hybridization. Intron 1 was relatively long (3.4 kbp) and was found to be inserted after the ninth codon (TAC) from the initiation codon. This position proved to be almost homologous to the conserved first intron insertion position in the vertebrate galectin genes (i.e. genes of mammalian galectin-1, -2, and -3 and chick 14-kDa galectin). On the other hand, intron 2 was much shorter (0.6 kbp), and it was inserted into the central region of the second carbohydrate-binding site. Although such an insertion pattern has never been observed in the vertebrate galectin genes, it seems to be common in C. elegans tandem repeat-type galectin genes, as predicted by the C. elegans genome project (Coulson, A., and the C. elegans Genome Consortium (1996) Biochem. Soc. Trans. 24, 289–291). Based on extensive sequence comparison, the origin and molecular evolution of the tandem repeat-type galectins are discussed.

Galactins are a group of animal lectins characterized by metal independence and specificity for β-galactosides (Refs. 1–7; for a recent collection of galectin topics, see Ref. 7). Although biological phenomena in which galectins are involved are extremely diverse, all members of the galectin family must meet two criteria: specific affinity for β-galactosides and an evolutionarily conserved sequence motif in the carbohydrate-binding site (1).

Galectins had been found only in vertebrates until we found a 32-kDa lectin that has a primary structure homologous to those of vertebrate galectins in the nematode Caenorhabditis elegans (8, 9). This worm was considered to be an excellent model animal (10–12) in which to study the genetics of galectins. Since we recently found that C. elegans also has a prototype galectin (13), it has become evident that the nematode has a set of galectins similar to that of mammals. Therefore, galectins may have been playing fundamental roles in the animal kingdom since the Precambrian era, conserving their fundamental structures and sugar binding specificity. This view is supported by the finding that a primitive metazoan, the marine sponge Geodia cydonium (14), also has galectins. More recently, galectins were also found in fungus (15).

The 32-kDa galectin of C. elegans was the first example not only of an invertebrate galectin, but also of a “tandem repeat-type” galectin. It is composed of two domains, each of which is homologous to typical vertebrate 14-kDa type galectins (~25% amino acid identity). Subsequently, it was suggested that all known galectins can be classified into three types in terms of molecular architecture, i.e. prototype, chimera-type, and tandem-repeat type (3, 4). In the case of mammals, galectin-1 (16–20), galectin-2 (21), galectin-5 (22), galectin-7 (23, 24), and Charcot-Leyden crystal (CLC) protein (or galectin-10) (25) can be classified as prototype; galectin-3 (26–30) as chimera-type; and galectin-4 (31, 32), galectin-6 (33), galectin-8 (34), galectin-9 (35), and probably PCTA-1 (or galectin-11) (36) as tandem-repeat type. Interestingly, the number of tandem-repeat type galectins found has recently been increasing. Although neither the biological functions nor the overall three-dimensional structures of tandem-repeat type galectins have been established, we recently proposed a possible role as a “heterobifunctional cross-linker” (37). This idea is supported by the observation that the first and second domains of rat galectin-4 show significantly distinct saccharide specificities (31). Recent immunohistochemical studies showed that the 32-kDa galectin is localized most abundantly in the adult cuticle of C. elegans (38), implying that it may be an essential component of the adult cuticular matrix for the construction of the tough and durable outer barrier of the worm’s body.

It is of interest to examine whether or not nematode and mammalian tandem-repeat type galectins originated from a common ancestor molecule. Here, we report the genomic structure of the 32-kDa galectin of the nematode C. elegans. By comparing the elucidated intron-exon organization with those of mammalian galectin genes identified so far (and also with those of several tandem-repeat type galectin genes in C. elegans identified by the recent genome project (44)), we may find...
a clue to the molecular evolution of the tandem repeat-type galectin genes. The fact that tandem repeat-type galectins are commonly found in nematodes (protostomes) and mammals (deuterostomes) implies that these cross-linking molecules are of fundamental importance.

**EXPERIMENTAL PROCEDURES**

**Cultivation of C. elegans and Preparation of Genomic DNA from C. elegans**—The nematode *C. elegans* was cultivated in liquid medium as described previously (13). The worms were harvested by conventional methods, and the genomic DNA was prepared as described (11).

**Physical Mapping of the 32-kDa Galectin Gene**—YAC1-gridded polynucleotide and Supply filters, YAC clone Y61H10, and the cosmid clones B0553, C13F7, C15B1, C18B5, F31B9, and W06D10 were provided by Dr. Alan Coulson (Sanger Center, Cambridge, United Kingdom). The filters were screened by hybridization using the α-32P-labeled probe as described previously (12). The DNA from Y61H10 and cosmids was prepared as described (12).

**Southern Hybridization Analysis**—The genomic DNA isolated from the nematode *C. elegans* or DNA isolated from YAC Y61H10 was digested with several restriction enzymes and subjected to 1% agarose gel electrophoresis. The DNA fragments were transferred to a nylon membrane filter (Hybond-N, Amersham Life Science, Inc.) in an alkaline solution according to the manufacturer's instructions. A fragment (0.8 kbp) of the 32-kDa galectin cDNA amplified by PCR was labeled using the Megaprime™ DNA labeling system (Amersham Life Science, Inc.) with [α-32P]dCTP (110 TBq/mmol). For probing introns 1 and 2, the products isolated by agarose gel electrophoresis were purified using Prep-A-Gene® DNA purification systems (Bio-Rad) to remove unincorporated dNTPs and excess primers. Purified intron fragments were also α-32P-labeled using the Megaprime™ DNA labeling system. Southern hybridization was performed essentially as described (39), except that Rapid-hyb buffer (Amersham Corp.) was used both for prehybridization and hybridization.

**Primers and Taq Polymerase Used for Polymerase Chain Reaction and DNA Sequencing**—Oligonucleotide primers were distinguished by the position number in the cDNA sequence (position 1 corresponds to A of the ATG translation initiation codon) (9). The letters “F” and “R” stand for “forward primer” and “reverse primer,” respectively. These primers were synthesized with an Applied Biosystems Model 392A DNA synthesizer. The sequences and positions of the primers are shown in Fig. 1. LA Taq (Takara Shuzo Corp.), which has a proofreading activity, was used for the polymerase chain reaction.

Plasmids for DNA sequencing were prepared using the Wizard™ Plus miniprep DNA purification system (Promega) according to the manufacturer's instructions. DNA sequences were determined by the dye termination cycle sequencing method using an ABI PRISM™ Dye Terminator Sequencing FS Ready Reaction kit with an Applied Biosystems Model 373S DNA sequencer according to the manufacturer's instructions.

**Amplification of the Two Introns by PCR—Oligonucleotide primers were designed to amplify fragments of the 32-kDa galectin gene using cDNA (9) and genomic DNA as templates. The PCR products obtained using genomic DNA were ligated to pCR1 (Invitrogen). For intron 1, primers −14F and 273R were designed to generate a PCR product, which, in turn, served as a template for an additional round of PCR using primers −5F and 17IR. For intron 2, primers 574F and 654R were used. These vectors were used to transform competent INVPlC cells, and the cloned DNAs were sequenced. Since, for an unknown reason, it was not possible to read the mostly internal sequences of intron 2 with primers used for the PCR product, the pCR1 vector, a secondary PCR product obtained using the 1.05-kbp PCR product (primers 422F and 840R) as a template and two internal primers (INT-F (5'-TTGGTGACAGTTATGCTG-3') and INT-R (5'-GTGGTGACAGGGCCCA-3')) was sequenced.

**Cloning of the Genomic DNA Fragments Coding the 32-kDa Galectin and Sequencing**—For cloning the genomic fragment including the 5'-upstream region, the coding region between the 5'-untranslated region and intron 1, and a part of intron 1, the genomic DNA was cleaved at the NdeI site in the upstream region predicted by Southern analysis and at the PsI site in intron 1. For cloning the genomic fragment from the 3'-end region of intron 1 to the 3'-noncoding region of the galectin gene, the fragment in pBluescript was digested with either Sall or PstI and cloned into the 5'-region or the 3'-region of the fragment, respectively. After digestion of the genomic DNA, DNA fragments were subjected to 1% agarose gel electrophoresis (Sea Kem® GTG® agarose, FMC Corp. BioProducts) and visualized in the presence of ethidium bromide, and then the gel blocks containing the desired bands were cut out with a razor blade. DNA was extracted using Prep-A-Gene® DNA purification systems. After ligation to the plasmid vector pBluescript, the vector was used to transform competent JM109 cells. The genomic library composed of Sall/PstI fragments of ~1.5–3 kbp was screened with a probe prepared by PCR using primer −5F and a reverse primer corresponding to the site just upstream of the PsI site in intron 1 (5'-CTGCAAGAGCAATTTGAAT-TCG-3'). The genomic library composed of Xbal/XhoI fragments of ~1.5–4 kbp was screened with an intron 2 probe. Colony hybridization was performed using Rapid-hyb buffer both for prehybridization and hybridization. After two rounds of screening, a single clone was isolated, and the DNA sequence was determined.

For DNA sequencing the 1.8-kbp Sall/PstI fragment in pBluescript, a Kilo-Sequence Deletion kit (Takara Shuzo Corp.) was used according to the manufacturer's instructions to generate deletion mutants. Briefly, the fragment in pBluescript was digested with either Sall plus KpnI or SpeI plus SacI from the 5'-region or the 3'-region of the fragment, respectively. After treatment with exonuclease III for varying lengths of time, the single-stranded part of the digested DNA was removed by mung bean nuclease treatment, and blunt-ending was conducted using the Klenow fragment. The pBluescript plasmid with blunt ends on both sides was recircularized with DNA ligase and used for transformation of competent JM109 cells. Approximately 50 colonies

---

1 The abbreviations used are: YAC, yeast artificial chromosome; kbp, kilobase pair(s); kb, kilobase(s); bp, base pair(s); PCR, polymerase chain reaction.
were picked up, and the plasmid was isolated from each colony using the Wizard™ Plus miniprep DNA purification system. Several colonies with inserts of varying length were chosen and subjected to DNA sequencing using the vector-specific T7 promoter primer (5'-TAATACGACTCACTATAGGG-3') and the T3 promoter primer (5'-CAATTACACCCCTGCTTAAAGG-3').

For DNA sequencing the 2.2-kbp XbaI/XhoI genomic fragment in pBluescript, inner primers (151F, 361F, 554F, 818F, 1006F, 1036F, 171R, 239R, 486R, 676R, 840R, 903R, and 1071R) corresponding to the coding region of the 32-kDa galectin, an intron 2-specific forward primer (5'-TGCCATGTTGTCATTTCTATTTCG-3'), and vector-specific primers (T7 and T3 promoter primers) were used for sequence walking.

**Northern Hybridization Analysis**—A cDNA fragment was labeled with [32P]dCTP and used as a probe for transcripts of the nematode 32-kDa galectin. Total RNA or poly(A)+ RNA from *C. elegans* (11, 12) was electrophoresed under denaturing conditions, transferred to a nylon membrane, and hybridized with the probe as described (39).

**PCR for the Detection of a trans-Spliced Leader Sequence**—A ZAP cDNA library of *C. elegans* (mixed stages) (provided by Dr. R. H. Waterston, Washington University School of Medicine) or cDNA obtained from *C. elegans* was used as a template for amplification of the 5'-end of the 32-kDa galectin cDNA by PCR. Two specific primers for two kinds of trans-spliced leader sequences (SL1 and SL2, 5'-GTTTAATTACCAAGTTGAGG' and 5'-GTTTTTAACCCAGTTACTCAAGG-3', respectively) were used as forward primers, and primer 239R was used as the reverse primer. The derived product of this primary PCR was diluted 100-fold and used as the template for the secondary PCR, in which the SL1 primer plus primer 214R or the SL2 primer plus primer 214R were used as primers. The main product (~370 bp) obtained with the SL1/214R primer set was used as the template for the third PCR using the SL1 primer and primer 171R. The product was inserted into the TA cloning vector pCRII, which was used to transform competent INVaF cells. The DNA sequence of the inserted fragment was determined.

**RESULTS**

**Physical Mapping of the 32-kDa Galectin Gene**—Using several combinations of primers corresponding to the coding region of the 32-kDa galectin and comparing the PCR product size of the reactions using cDNA and genomic DNA as templates, we found that an intron (which turned out to be intron 2 described below) exists in the coding region of the second lectin domain. A 1.05-kbp PCR product, obtained using primers 422F and 840R, contained intron 2 and parts of the neighboring exons. This fragment was used to screen polytene filters to locate the 32-kDa galectin gene in the genome of *C. elegans*. YAC polytene filters consist of a grid of 958 YAC clones (12, 40), which were selected to give a 2-fold coverage of the whole genome of *C. elegans*. YAC clone Y61H10 was found to give a positive hybridization signal (Fig. 2A).

We isolated DNA from cosmid clones that overlap YAC Y61H10 (B0553, C13F7, C15B1, C18B8, F31B9, and W06D10, provided by the Sanger Center) and conducted PCR with several combinations of primers. However, none of them gave a positive product. A supplementary (Supphy) grid of 223 YAC clones (also provided by the Sanger Center) was screened with the intron 2 probe. Clones Y81G3 and Y48C3, which overlap with Y61H10, were found to be positive (data not shown). When the DNA isolated from YAC Y61H10 was digested with various restriction enzymes and subjected to Southern blot analysis with the cDNA probe (Fig. 2C), the size of each positive hybridization signal matched those obtained by Southern blot analysis of genomic DNA isolated from the worm (Fig. 2B). We conducted PCR using Y61H10 DNA as a template with primers 422F and 840R and obtained only one product, the size of which...
**Structure of the 32-kDa Galectin Gene of *C. elegans***

| Primer combination | Template | cDNA (theory) | Genome (observed) | Estimated intron size, Δ (genome – cDNA) |
|--------------------|----------|---------------|-------------------|----------------------------------------|
| 14F 171R           | 14F 171R | 2.5 kb        | 2.3 kb (intron 1) |                                        |
| 14F 273R           | 289 b    | 2.6 kb        | 2.3 kb (intron 1) |                                        |
| 5F 171R            | 177 b    | 2.5 kb        | 2.3 kb (intron 1) |                                        |
| 5F 273R            | 289 b    | 2.6 kb        | 2.3 kb (intron 1) |                                        |
| 5F 437R            | 443 b    | 2.9 kb        | 2.3 kb (intron 1) |                                        |
| 15F1 273R          | 124 b    | 0.12 kb       | 0                 |                                        |
| 15F1 437R          | 287 b    | 0.28 kb       | 0                 |                                        |
| 361F 486R          | 126 b    | 0.12 kb       | 0                 |                                        |
| 422F 654R          | 233 b    | 0.85 kb       | -0.62 kb (intron 2)|                                        |
| 422F 840R          | 419 b    | 1.05 kb       | -0.62 kb (intron 2)|                                        |
| 554F 676R          | 123 b    | 0.75 kb       | -0.62 kb (intron 2)|                                        |
| 554F 840R          | 287 b    | 0.92 kb       | -0.63 kb (intron 2)|                                        |
| 574F 654R          | 81 b     | 0.70 kb       | -0.62 kb (intron 2)|                                        |
| 818F 1071R         | 254 b    | 0.25 kb       | 0                 |                                        |

* a, b, base.

**Table 1**

**Summary of PCR analysis of the 32-kDa *C. elegans* galectin gene**

(1.05 kb) was the same as that produced using the genomic DNA as a template. These results confirmed that the 32-kDa galectin gene is mapped to Y61H10 and thus is located on chromosome II (where Y61H10 is mapped).

**Identification of the Two Introns by Genomic PCR**—Since it is important to know the sizes and insertion points of introns to elucidate the structure of the 32-kDa galectin gene, the sizes of the PCR products obtained using cDNA and genomic DNA as templates were compared. The primers used are indicated in Fig. 1. First, the region connecting the two lectin domains was searched. Primers 361F and 486R were used, but no evidence for the existence of an intron was obtained. Then, >20 combinations of primers covering the whole coding sequence were used. When the sizes of the two reaction products were different, inner primers were used for the next PCRs to narrow the location of the intron. The results, summarized in Table 1, suggested the existence of at least two introns: one between primers 5F and 171R and the other between primers 574F and 840R (Fig. 1). Amplification of the 32-kDa galectin gene is mapped to Y61H10 and thus is located on chromosome II (where Y61H10 is mapped).

**Sequence of the Two Introns**—The sequences of the above two PCR products inserted into the pCRII vector were determined by dideoxy cycle sequencing. For the 5F/171R product, the vector-specific T7 promoter primer and the M13 reverse primer were used. Although the insert was not long (~0.6 kb), determination of the whole sequence by one reaction in either the forward or reverse direction was not possible. Fluorescence signals on the electropherogram always weakened significantly at a certain position and became undetectable beyond this position, probably due to the secondary structure of the template. Therefore, two intron-specific inner primers, INT-F (5'-TGCTCGAGCT-TATGGCGT-3') and INT-R (5'-GTTTTGAGAAGCCGCA-3'), were synthesized to amplify the inner part. The product of the secondary PCR was inserted into the pCRII vector, and its whole size was determined. It appeared that a single intron (intron 2) was inserted between G (603th nucleotide of cDNA) and C (604th nucleotide). This region corresponds to the central part of the second lectin domain (Fig. 3, intron 2). The whole sequence of intron 2 will appear in the DDBJ, EMBL, and GenBank™ nucleotide sequence data bases with accession number D85887.

**Genomic Cloning of Fragments Containing the 5'- and 3'-Noncoding Regions**—Since introns in *C. elegans* genes are generally smaller than those of other eukaryotes (41), some small introns might have been overlooked by the present procedure, which is based on comparison of the PCR products of the genomic DNA and cDNA. Moreover, the primers used corresponded only to the coding region. Therefore, if there were introns in the 5'- and 3'-noncoding regions, they would not be detected. We tried to isolate genomic clones containing the 32-kDa galectin gene from the phage genomic library. However, no positive clone was obtained. A possible explanation of this is that the intron 2 region contains some specific secondary structures that make it resistant to cloning into the phage vector of the genomic library. Thus, we tried to create a library of restriction enzyme fragments and to obtain fragments containing the noncoding region by colony hybridization. To find restriction enzyme sites in these regions favorable for cloning, genomic Southern analysis was conducted with an α-32P-labeled 0.8-kb cDNA fragment including the whole coding region (Fig. 2B). The result suggested that the 32-kDa galectin is encoded by a single copy gene. When the genomic DNA was digested with *SalI* and hybridized with the cDNA probe, two bands were detected (Fig. 2B). This indicated that the coding region is divided into two fragments by the *SalI* site. When the intron 1 fragment was labeled and used as a probe, only the band of ~4 kb was detected (data not shown). This fragment seemed to include the 5'-part of the coding region and the 5'-noncoding region. Since there is a *SalI* site in the coding region of the 32-kDa galectin (Fig. 3A), there should be a *SalI* site at ~1.2 kb upstream from the ATG translation initiation site. For cloning the upstream genomic region, this *SalI* site and the *PstI* site previously found in intron 1 were chosen (the estimated size of this *SalI/PstI* fragment is 1.5–2 kb). The whole DNA sequence of this fragment was determined by making deletion mutants of various lengths. The 3'-part was found to match precisely the cDNA sequence plus intron 1. This means that no other intron exists upstream of intron 1.

When the genomic DNA was digested with *XhoI*, a fragment of ~6 kb was detected (Fig. 2B). Since there is no *XhoI* site either in the coding region or in the two already identified introns, it seemed that there were two *XhoI* sites, one in the upstream region and the other in the downstream region of the 32-kDa galectin gene. Therefore, the *XhoI* site was chosen for the 2.5-kb insert (2.5 kbp) was too big for the conventional polymerase. We inserted the PCR products obtained with primers 361F and 486R into the vector-specific T7 promoter primer and the M13 reverse primer. The product of the secondary PCR was inserted into the pCRII vector, and its whole size was determined. It appeared that a single intron (intron 2) was inserted between G (603th nucleotide of cDNA) and C (604th nucleotide). This region corresponds to the central part of the second lectin domain (Fig. 3, intron 2). The whole sequence of intron 2 will appear in the DDBJ, EMBL, and GenBank™ nucleotide sequence data bases with accession number D85887.
the 3' -end of the desired fragment. For the 5' -end, the Xba I site in intron 1 was chosen. The predicted size of this Xba I/Xho I fragment was ~2–2.5 kbp. The same procedure that had been applied to the isolation of the Sal I/Pst I fragment was used. The cloned Xba I/Xho I fragment was 2.2 kbp long, as expected. It contained the 3' -end part of intron 1 from the Xba I site, the exon matching the cDNA sequence (from nucleotides 28 to 603), intron 2, and another exon matching the cDNA sequence (from nucleotides 604 to 1071 including the 3' -noncoding sequence), except that the repeat number of C in the 3' -noncoding region was 17 instead of 15 in cDNA. This difference is probably due to the difficulty in sequencing the homopolymer region by means of dye terminator reactions. The absence of other introns in the genomic regions between introns 1 and 2 and downstream of intron 2 was confirmed.

The lengths of introns 1 and 2 were 2429 and 631 bp, respectively. The sequences of the intron-exon boundaries are indicated in Fig. 4. The donor and acceptor splice sites in mRNA present in both introns conform to the GT-AG rule (42).

**Transcript Analysis**—The cDNA clone previously isolated (9) had a rather short 5' -untranslated region (only 14 nucleotides upstream of ATG). This short untranslated region was found to match the genomic sequence. To investigate further the untranslated region, the attachment of the trans-spliced leader sequence to the 5' -end of the galectin gene transcript was checked. Since a large fraction of the messages in *C. elegans* is known to be trans-spliced to the SL1 leader sequence (22 nucleotides; see Fig. 4) (12, 43), the galectin mRNA was amplified by reverse transcription-PCR using the SL1-specific primer and a consecutive series of reverse primers (239R, 214R, and 171R), similar to the 5' -rapid amplification of cDNA ends method. Production of a specific fragment was observed for the third PCR (since the forward primers were always the same, three consecutive reactions were done using these reverse primers to better specify the product); this fragment was inserted into the pCR II vector and the sequence was determined. The amplified product contained the SL1 sequence (GGTTTTAATTCCAAAGTTTGAG) followed by GAACCACCGAAAAGCCACCACCAAGCAGAGTG, which corresponds to the genomic sequence of the galectin (the underlined ATG is the translation initiation codon) (Fig. 4). This showed that the trans-spliced leader sequence SL1 was added to the transcript of the 32-kDa galectin at position ~30 (Fig. 4) and that no signal sequence precedes the ATG translation initiation codon. Another primer specific to the trans-spliced leader sequence SL2 (GGTTTTAGACCCAGATTACTCAAG) was also examined, but no product was detected. Therefore, the SL1 trans-spliced leader sequence seemed to be added specifically to the 32-kDa galectin transcript.

**Northern Blot Hybridization Analysis**—The total RNA and mRNA fraction were analyzed by hybridization reaction with the cDNA probe labeled with 32P. Hybridization was seen for both fractions at the position corresponding to the size of ~1.4 kb (Fig. 5). The expected size of mRNA for the 32-kDa galectin excluding poly(A) is 1123 nucleotides (22 for SL1, 30 for the 5'-untranslated region, 840 for the coding region, and 231 for the 3'-untranslated region). The mRNA length estimated from Fig. 5 suggests that the poly(A) length is ~300 nucleotides.

**Summary of Analysis of the 32-kDa Galectin Gene**—The 32-kDa galectin gene of the nematode *C. elegans* is located on YAC Y61H10 on chromosome II. It seems to be a single copy gene according to the results of genomic Southern blotting analysis. Two introns, intron 1 (2429 bp) and intron 2 (631 bp), were identified and cloned by genomic PCR, and the DNA sequences were determined. Intron 1 is inserted near the 5'-end of the coding region, and intron 2 at the center of the coding region of the second lectin domain. The gene fragment including the 5'-noncoding region and the 5'-part of intron 1 was isolated as a Sal I/Pst I fragment (2.2 kbp). The fragment including the 3'-part of intron 1, intron 2, and the 3'-noncoding region was isolated as an Xba I/Xho I fragment (1.8 kbp). Therefore, the whole genomic sequence of the 32-kDa galectin gene was determined. The entire nucleotide sequence (5535 bp) of the 32-kDa galectin gene from the Sal I site to the Xho site will appear...

---

*Structure of the 32-kDa Galectin Gene of C. elegans*
in the DDBJ, EMBL, and GenBank™ nucleotide sequence data bases with accession number AB000802. The analysis of the transcript showed that SL1 was trans-spliced at position −30, and the sequence AATAAA found at the end of the cDNA seemed to be the poly(A) signal. The size of the transcript was estimated to be 1.4 kb by Northern blot analysis.

**DISCUSSION**

By screening polytene and Suppoly filters derived from *C. elegans*, the gene encoding the 32-kDa galectin was located on YAC Y61H10, which has been mapped by chromosome II of the organism. However, six overlapping cosmids did not include the 32-kDa galectin gene. Recently, the cosmid W09H1 was sequenced by the *C. elegans* genome project (44) and shown to contain the 3′-end part of the galectin gene, i.e., downstream of intron 2, including −300 bp of intron 2. Although W09H1 does not overlap with YAC Y61H10 according to the available physical map of *C. elegans*, this could be due to a false-negative finding of poor hybridization of Y61H10 with W09H1 on the grid of cosmid clones.

By Northern blot analysis, a transcript of ≈1.4 kb was detected even though the amount of total RNA loaded on the gel was only 1.5 μg (Fig. 5). The transcript of this galectin should be relatively abundant, and this is consistent with the result of purification of 32-kDa galectin protein, which was obtained in a yield of ≈250 μg of protein from 5 g of worms (wet weight) (8). This yield must be lower than the real amount present in the nematode since some of the protein remains in the lactose-insoluble fraction (38). The mRNA for the 32-kDa galectin was found to contain the trans-spliced leader sequence SL1 at the 5′-end. trans-Splicing, which is an interesting phenomenon observed in *C. elegans* (12) and also in trypanosomes (45), is defined as the ligation of exon sequences originating from independently transcribed RNAs, although its biological meaning is unknown. SL1 is the most common trans-spliced leader sequence found in *C. elegans* mRNAs (43). More than 40% of the messages in *C. elegans* are trans-spliced to the SL1 leader sequence (46). The 32-kDa galectin gene seems to fulfill the requirements for a recipient of the SL leader sequence, i.e., having a 3′-splice AG acceptor, but lacking a 5′-splice donor sequence in the 5′-untranslated region of the gene (47, 48). Comparison of the sequences of cDNA and genomic DNA suggests that the AG dinucleotide sequence at positions −32 and −31 (underlined in Fig. 4) is the acceptor signal. Since SL1 was trans-spliced to the 5′-region of the galectin gene transcript, it was difficult to determine the start site of transcription. The spliced-out sequence (called an "outron") that initially extended from the 5′-end of the transcript no longer exists and thus is not

---

2 A. Coulson, personal communication.
detectable. The sequence AATAAA found downstream of the TAA termination codon proved to be the poly(A) signal because the poly(A) sequence appeared after 13 base pairs (Fig. 4). In *C. elegans*, polyadenylation usually begins ~13 nucleotides downstream of the AAUAAA consensus sequence (12).

The intron-exon architecture of the 32-kDa galectin has two noteworthy characteristics. The first is the absence of an intron between the two lectin domains. The presence of multiple lectin domains in the same polypeptide chain was reported in the case of the macrophage mannose receptor, which has multiple C-type lectin domains (49, 50). These domains are considered to have been formed by gene duplication from the original ances-
tral gene. Introns exist at all (51, 52) or most (53) of the sites connecting the coding regions corresponding to each lectin domain. Although the present 32-kDa galectin gene should have been formed by gene duplication, no intron was found between the two regions. One possible explanation is that the noncoding structure (intron) that had originally existed was lost after gene duplication. The process of gene duplication might have been different between tandem repeat-type galectins and mammalian macrophage mannose receptors.

The second characteristic of the genomic structure of the 32-kDa galectin is its unique insertion pattern of introns. Fig. 6 shows the structures of galectin genes so far reported: human galectin-1 (54) and galectin-2 (21), mouse galectin-1 (56) and galectin-3 (57), and chicken 14-kDa galectin (58). All genes for the prototype galectins (mammalian galectin-1 and chicken 14-kDa galectin, which have a molecular size of 14–16 kDa and exist mainly as a dimer in solution) and the galectin domain of the chimera-type galectin (mouse galectin-3) are interrupted with three introns in a conserved manner. The coding region of these galectins consists of four exons, from the 5′-end to the 3′-end: exons I (6–9 bp), II (80–83 bp), III (160–172 bp), and IV (144–150 bp). The largest, exon III, encodes the most conserved, hydrophilic region, including the amino acids considered to be important in forming hydrogen bonds or in van der Waals interaction with the bound sugar (indicated at the top of Fig. 6) (59–63). Intron-exon organization is very different in the case of the C. elegans 32-kDa galectin gene. Only intron 1 was found to be conserved between C. elegans and mammals, although it was very much longer than the corresponding intron in mammalian galectin genes. Intron 2 was found to interrupt the central part of the second lectin domain. As a result, the region coding a conserved hexapeptide (His-Phe-Asn-Pro-Arg-Phe, with residues required for sugar binding underlined) was separated from the remaining part including Ile-Arg-Asn-Ser-Leu-Ala-Ala-Asn-Glu-Trp-Gly-Asn-Glu-Glu-Arg. Such a type of insertion has never been observed in mammalian galectin genes. It would be interesting to know whether or not such a unique intron-exon organization is general in tandem repeat-type galectin genes, such as the mammalian galectin-4 and -8 genes.

Recently, the genome project on C. elegans (44) suggested that multiple genes encode proteins having the characteristic sequence motif of the galectin family (originally found by Dr. D. N. W. Cooper) (7). Among them, at least five (including the 32-kDa galectin) fall into the category of tandem repeat-type galectins according to systematic predictions made using the Genefinder program. The predicted amino acid sequences are aligned with the 32-kDa galectin sequence in Fig. 7. The predicted positions of introns (not necessarily confirmed) were found to be widely conserved in either of the two domains of these tandem repeat-type galectin genes. Therefore, the special intron insertion pattern revealed by this work seems to be fairly common in the galectin gene family of C. elegans. Work on these galectin genes in C. elegans will be described elsewhere.4

When the amino acid sequences of the two domains of individual tandem repeat-type galectins are compared, the observed percent amino acid identities are similar (30–36%) in mammals (galectin-4, -8, and -9) and C. elegans (galectins encoded by cosmids F52H13, ZK892, C44F1, and ZK1248). This fact suggests that a gene duplication event could have occurred at almost the same time in both deuterostomes and protostomes. On the other hand, the observed percent amino acid identities among the first domains of mammalian and C. elegans tandem repeat-type galectins are relatively low (21–29%) compared with those among the second domains (36–41%). One possible explanation for this discrepancy is different molecular evolutionary rates in the first and the second domains. If this is the case, the two domains should have acquired distinct, although basically similar (i.e. galactose-specific) carbohydrate binding properties in the course of molecular evolution. In fact, we have recently shown that the binding affinity of the expressed first domain (N-terminal half) for asialofetuin is much weaker than that of the second domain (C-terminal half) in the 32-kDa galectin of C. elegans (37). As regards fine sugar binding specificity, it was also reported that the first and second domains of rat galectin-4 showed preferential specificity for Galβ1–4GlcnAc and Galβ1–3GalNAc, respectively. Further studies are needed to clarify the divergence of biological functions of individual domains of these tandem repeat-type galectins in both nematodes and mammals.

There must be a special reason for the maintenance of so many tandem repeat-type galectins in both nematodes and mammals. One possibility is the valency. We previously proposed that these molecules act as heterobifunctional linkers (37). Simple saccharides such as lactose are not effective in extracting either the nematode 32-kDa galectin (38) or galectin-8 (34), as would be expected if these galectins bind to the matrix via dual specific interactions, which would not be possible for the homobifunctional prototype galectins. A similar suggestion was made for galectin-4 (31). Future research on C. elegans galectins is expected to contribute greatly to our understanding of not only galectin-carbohydrate interaction, but also the physiological importance of the tandem repeat-type carbohydrate-binding proteins, including the 32-kDa galectin.

Acknowledgments—We are grateful to the Caenorhabditis Genetics Center (which is supported by the National Institutes of Health National Center for Research Resources) for providing wild-type C. elegans. We are also grateful to Dr. Alan Coulson for providing the polytene and Suppol Y AC filters, Y AC clone Y61H10, and the six cosmid clones, and we thank Dr. R. H. Waterston for providing the ZAP cDNA library. We thank Dr. D. N. W. Cooper for information on homologous galectins in C. elegans found by BLAST programs (55) and for helpful discussions. Thanks are also due to M. Sekiguchi, T. Ubukata, Y. Kobayashi, S. Hayakawa, and H. Terada for technical assistance.

REFERENCES
1. Barondes, S. H., Castronovo, V., Cooper, D. N. W., Cummings, R. D., Drickamer, K., Feizi, T., Gitt, M. A., Hirabayashi, J., Hughes, C., Kasai, K., Leffler, H., Liu, F.-T., Lazebnik, Y., McCaffery, J. M., Melancon, B. R., Massaro, D. (1988) Biochemistry 27, 692–699
2. Barondes, S. H., Cooper, D. N. W., Gitt, M. A., and Leffler, H. (1994) J. Biol. Chem. 269, 20807–20810
3. Barondes, S. H., Cooper, D. N. W., Gitt, M. A., and Leffler, H. (1994) J. Biol. Chem. 269, 20807–20810
4. Kasai, K., and Hirabayashi, J. (1996) J. Biochem. (Tokyo) 111, 1–8
5. Kasai, K., and Hirabayashi, J. (1996) J. Biochem. (Tokyo) 111, 553–555
6. Hirabayashi, J., Satoh, M., Ohyama, Y., and Kasai, K. (1992) J. Biochem. (Tokyo) 111, 553–555
7. Hirabayashi, J., Satoh, M., and Kasai, K. (1992) J. Biochem. (Tokyo) 111, 553–555
8. Hirabayashi, J., Satoh, M., and Kasai, K. (1992) J. Biochem. (Tokyo) 111, 553–555
9. Hirabayashi, J., Satoh, M., and Kasai, K. (1992) J. Biochem. (Tokyo) 111, 553–555
10. Brenner, S. (1974) Genetics 77, 71–84
11. Wood, W. B. (1988) The Nematode Caenorhabditis elegans, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY
12. Epstein, H. F., and Shakes, D. C. (eds) (1985) Methods Cell Biol. 48
13. Hirabayashi, J., Ubukata, T., and Kasai, K. (1996) J. Biol. Chem. 271, 2497–2505
14. Pfeifer, K., Haasemann, M., Gasmin, V., Bretting, H., Fahrenholz, F., and Muller, W. G. (1993) Glycobiology 3, 179–184
15. Cooper, D. N. W., Boulianne, R. P., Charlton, S., Farrell, E. M., Sucher, A., and Lu, B. C. (1997) J. Biol. Chem. 272, 1514–1521
16. Hirabayashi, J. (1988) J. Biochem. (Tokyo) 104, 1–8
17. Hirabayashi, J., Ayaki, H., Soma, G., and Kasai, K. (1989) Biochim. Biophys. Acta 1008, 85–91
18. Clerch, L. B., Whitney, P., Hass, M., Brew, K., Miller, T., Werner, R., and Massaro, D. (1988) J. Biol. Chem. 263, 2597–2598
19. Wells, V., and Mallucci, L. (1991) Cell 64, 91–97
20. Abbott, W. M., Mellor, A., Edwards, Y., and Feizi, T. (1989) Biochem. J. 259, 293–296
Structure of the 32-kDa Galectin Gene of C. elegans

21. Gitt, M. A., Massa, S. M., Leffler, H., and Barondes, S. H. (1992) J. Biol. Chem. 267, 10601–10606
22. Gitt, M. A., Wiser, M. F., Leffler, H., Herrmann, J., Xia, Y.-R., Massa, S. M., Cooper, D. N. S., Luís, A. J., and Barondes, S. H. (1995) J. Biol. Chem. 270, 5032–5038
23. Madsen, P., Rasmussen, H. H., Flint, T., Grumov, P., Kruse, T. A., Honore, B., Vorum, H., and Celis, J. E. (1995) J. Biol. Chem. 270, 5823–5829
24. Magnaldo, T., Bernerd, F., and Darmon, M. (1995) Dev. Biol. 168, 259–271
25. Ackerman, S. J., Corrette, S. E., Rosenberg, H. F., Bennett, J. C., Mastroianni, D. M., Nicholson-Weber, A., Weller, P. F., Chin, D. T., and Tennen, D. G. (1993) J. Immunol. 150, 456–468
26. Robertson, M. W., Albrant, K., Keller, D., and Liu, F.-T. (1990) Biochemistry 29, 8093–8100
27. Cherayil, B. J., Chaitovitz, S., Wong, C., and Pillai, S. (1990) Proc. Natl. Acad. Sci. U. S. A. 87, 7324–7328
28. Oda, Y., Leffler, H., Sakakura, Y., Kasai, K., and Barondes, S. H. (1991) Gene 99, 279–283
29. Albrant, K., Orida, N. K., and Liu, F.-T. (1987) Proc. Natl. Acad. Sci. U. S. A. 84, 6859–6863
30. Jia, S., and Wang, J. L. (1988) J. Biol. Chem. 263, 6009–6011
31. Oda, Y., Herrmann, J., Gitt, M. A., Turck, C. W., Burlingame, A. L., Barondes, S. H., and Leffler, H. (1993) J. Biol. Chem. 268, 5929–5939
32. Chiariotti, L., Wells, V., Bruni, C. B., and Mallucci, L. (1991) J. Biol. Chem. 266, 10912–10916
33. Gitt, M. A., Colnot, C., Xia, Y.-R., Atchison, R. E., Luís, A. J., Poirier, F., Barondes, S. H., and Leffler, H. (1995) Glycoconjugate J. 12, 548
34. Hadari, Y. R., Paz, R., Dekel, R., Mestrovic, T., Azciti, D., and Zick, Y. (1995) J. Biol. Chem. 270, 3447–3453
35. Sahin, U., Tureci, O., Schmitt, H., Cochevius, B., Johannes, T., Schmits, R., Luo, F. S. G., Schobert, I., and Pfreundschuh, M. (1995) Proc. Natl. Acad. Sci. U. S. A. 92, 11810–11815
36. Su, Z.-Z., Lin, J., Shen, R., Fisher, P. E., Goldstein, N. I., and Fisher, P. B. (1996) Proc. Natl. Acad. Sci. U. S. A. 93, 7252–7257
37. Arata, Y., Hirabayashi, J., and Kasai, K. (1995) J. Biol. Chem. (Tokyo) 121, 1002–1009
38. Arata, Y., Akimoto, Y., Hirabayashi, J., Kasai, K., and Hirano, H. (1996) EMBO J. 15, 201–207
39. Sambrook, J., Fritsch, E. F., and Maniatis, T. (1989) Molecular Cloning: A Laboratory Manual, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY
40. Coulson, A., Kuzono, Y., Lutterbach, B., Shownkeen, R., Sulston, J., and Waterston, R. (1991) Bioessays 13, 413–417
41. Blumenthal, T., and Thomas, T. (1988) Trends Genet. 4, 305–308
42. Breathnach, R., and Chambon, P. (1981) Biochem. Biophys. Res. Commun. 104, 173–177
43. Krause, M., and Hirsh, D. (1987) Cell 49, 753–761
44. Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990) J. Mol. Biol. 215, 403–410
45. Spieth, J., Brooke, G., Kuersten, S., Lea, K., and Blumenthal, T. (1995) Cell 73, 521–532
46. Cooper, D. N. W., Lusis, A. J., and Barondes, S. H. (1995) J. Biol. Chem. 270, 3447–3453
47. Harris, N., Super, M., Rits, M., Chang, G., and Ezekowitz, R. A. (1992) Blood 80, 2561–2573
48. Kim, S. J., Ruiz, N., Bezoisika, K., and Drickamer, K. (1992) Genomics 14, 721–727
49. Harris, N., Peters, L. L., Eicher, E. M., Rits, M., Raspberry, D., Kirkham, Q., Super, M., and Ezekowitz, R. A. (1994) Biochem. Biophys. Res. Commun. 198, 682–692
50. Bezoisika, K., Chishlow, G. V., Rose, J. M., Taylor, M. E., and Drickamer, K. (1991) J. Biol. Chem. 266, 11604–11609
51. Arata, Y., Hirabayashi, J., and Kasai, K. (1991) J. Biol. Chem. 266, 2363–2373
52. Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990) J. Mol. Biol. 215, 403–410
53. Spieth, J., Brooke, G., Kuersten, S., Lea, K., and Blumenthal, T. (1995) Mol. Cell. Biol. 15, 31770–31776
54. Spieth, J., Brooke, G., Kuersten, S., Lea, K., and Blumenthal, T. (1995) Mol. Cell. Biol. 15, 31770–31776
55. Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990) J. Mol. Biol. 215, 403–410
56. Chiariotti, L., Wells, Y., Bruni, C. B., and Mallucci, L. (1991) Biochim. Biophys. Acta 1089, 54–60
57. Chisso, G. C., Neri, S., and Liu, F.-T. (1992) Biochemistry 31, 9533–9538
58. Ohayma, Y., and Kasai, K. (1987) J. Biochem. (Tokyo) 104, 173–177
59. Hirabayashi, J., and Kasai, K. (1991) J. Biol. Chem. 266, 23645–23653
60. Hirabayashi, J., and Kasai, K. (1994) Glycoconjugate J. 11, 437–442
61. Liu, F.-T., Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990) J. Mol. Biol. 215, 403–410
62. Liao, D.-I., and Herzberg, O. (1994) Proc. Natl. Acad. Sci. U. S. A. 91, 1424–1432
63. Bourne, Y., Bolgiano, B., Liao, D.-I., Strecke, G., Cantau, P., Herzberg, O., Feizi, T., and Camillia, C. (1994) Nat. Struct. Biol. 1, 863–870