Accelerated Evolution in the Protein-coding Region of Galectin cDNAs, Congerin I and Congerin II, from Skin Mucus of Conger Eel (Conger myriaster)

Tomohisa OGAWA,*,† Chihiro ISHIH,*, Daiji KAGAWA,*, Koji MURAMOTO,* and Hisao KAMIYA**

*Department of Biological Resource Sciences, Graduate School of Agricultural Science, Tohoku University, Sendai 981-8555, Japan
**Department of Marine Biochemistry, School of Fisheries Sciences, Kitasato University, Sanriku, Iwate 022-0101, Japan

Received January 13, 1999; Accepted March 17, 1999

Two cDNAs encoding galectins named congerin I and II from the skin mucus of conger eel (Conger myriaster) were isolated and sequenced. Comparison of the nucleotide sequences of congerins I and II showed that the sequence similarities of the 5' and 3' untranslated regions (86 and 88%, respectively) were much higher than those of the protein-coding region (73%). The lengths of nucleotide substitutions per site (Ks) for the untranslated regions are smaller than the numbers of nucleotide substitutions per synonymous site (Ks) for the protein coding region. Furthermore, nonsynonymous nucleotide substitutions have accelerated more frequently than synonymous nucleotide substitutions in the protein-coding region (Ks/Ka = 2.57). These results suggest that accelerated substitutions have occurred in the protein-coding regions of galectin genes to generate diverse galectins with different molecular properties. Northern blot analysis showed that both congerins were expressed not only in the skin tissues but also in the stomach of conger eel.

Key words: galectin; skin mucus; conger eel; congerin I and congerin II; accelerated evolution

Galectin is a member of the lectin family of the members of which have at least one characteristic carbohydrate recognition domain (CRD) with an affinity for β-galactoside, and share the conserved sequence elements in the primary structure. To date, many kinds of galectins are known from vertebrates such as mammals, birds, amphibians, and invertebrates including nematodes and a sponge. They are classified into three types according to their primary structure: (1) the proto-type, which is composed of two identical 14–16 kDa subunits including only a single CRD domain per subunit; (2) the chimera type, which is composed of a CRD domain and an additional functional domain such as a heteronuclear ribonucleoprotein (hn RNP)-like domain; and (3) the tandem-repeat type, which is composed of two tandem similar CRD domains. Furthermore, galectins are widely distributed in various tissues and cells, and are involved in several biological phenomena through cell-cell interactions, cell adhesion, cell signalling, and cell growth.

Skin mucus of fishes is known to have many biologically active peptides and proteins, including hemolysins, lysozyme, and lectins, which are most likely related to defense against the invasion of pathogenic bacteria and parasites. Previously, we isolated two galectins named congerin I and congerin II from the skin mucus of conger eel (Conger myriaster), and analyzed their amino acid sequences. Both congerins I and II have homodimeric subunit structures with an N-acetylated Ser at the N-terminus and are classified as proto-type galectins. Congerins I and II have 136 and 135 amino acid residues, respectively, and share 48% amino acid sequence identity. Congerins I and II have different molecular properties such as thermostability, pH stability, and binding specificity for some lactose derivatives. For further elucidation of the structure-function relationships of galectins, the in vitro mutagenesis techniques using cloned cDNAs of two congerins appears to be the most promising approach.

In this paper, the cDNAs encoding congerin I and congerin II were cloned and their nucleotide sequences were analyzed. Comparison of the cDNA sequences for congerins I and II showed that the nucleotide sequences of the untranslated regions are much more conserved than those of the protein-coding regions. On the basis of this result, we discuss the molecular evolution of conger eel galectins.

Materials and Methods

Materials. Restriction endonucleases and other modifying enzymes were obtained from Takara Shuzo Co. (Kyoto, Japan) and Boehringer Mannheim (Germany). Synthetic oligonucleotides were obtained from Life Technologies (Tokyo, Japan). All other reagents were of analytical grade.

Construction of conger eel skin cDNA library. The skin of conger eel (Conger myriaster) from the Pacific coast of Miyagi prefecture, Japan, was sliced off after
freezing with liquid nitrogen, and then stored at −80°C. Total RNA was extracted by the acid guanidinium thiocyanate-phenol-chloroform extraction method with minor modifications. The poly(A) RNA fraction was purified on an oligo d(T) spin column (Pharmacia, Sweden), before cDNA synthesis (RPNI1256, Amersham Life Science, Cleveland, Ohio, USA) and ligated into the phage vector λ MOSElux (RPNI1712, Amersham Life Science). Infectious phage particles were obtained from the in vitro packaging (Gigapack III Plus, Stratagene, La Jolla, CA, USA) and were used to infect Escherichia coli ER1647 cells. The resulting library contained 4 × 10^6 independent phage clones.

Screening of congerin I and II clones from cDNA library. To obtain long and specific probes for screening, the polymerase chain reaction (PCR) was used with the specific oligonucleotide primers shown in Fig. 1 and the cDNA library as the template (30 cycles, 94°C for 1 min, 45°C or 50°C 1 min, 72°C 2 min). PCR products were subcloned into the Bluescript II SK + vector at the Sma I site and sequenced. After the sequences of PCR products were confirmed, digoxigenin-labeled probes were prepared by a DIG High Primer Labeling and Detection Starter kit I (Boehringer Mannheim Biochemica, Germany) according to the manufacturer’s instructions, and used for screening. The hybridized filters were washed successively twice at 65°C with 2× standard saline citrate (SSC)–0.1% SDS, and twice at 65°C with 0.1× SSC–0.1% SDS. The resulting positive clones were subcloned in the lysogenic bacterial strain BM 25.8 (RPNI1715, Amersham Life Science) to allow plasmid excision.

DNA sequencing and data analysis. Selected cDNA clones were sequenced in both directions using the T7 gene-10 and SP6 promoter primers, and internal primers (Fig. 1) by the dye deoxychain-termination method using a Thermo Sequenase dye terminator cycle sequencing pre-mix kit (Amersham Life Science) and an Applied Biosystems Model 373A DNA sequencer. The number of nucleotide substitutions per site (Ks) in the 5’- and 3’-untranslated regions (UTRs) and the numbers of nucleotide substitutions per synonymous site (Ks) and per non-synonymous site (Ka) for the protein-coding region between congerin I and II were computed by the method of Nei and Gojobori using the ODEN package developed by Y. Ina (National Institute for Genetics, Mishima, Japan). A synonymous site is a position in a codon at which a base substitution causes no amino acid change. A nonsynonymous site is a position in a codon at which a base substitution causes an amino acid change. The nucleotide sequence data reported in this paper are available from DDBJ, EMBL and Genbank databases with accession numbers of AB010276 and AB010277 for cDNAs encoding congerin I and congerin II, respectively.

Amino acid sequence analysis. The nucleotide sequencing of congerin I showed a discrepancy in the amino acid sequence at the 135th residue. Therefore, the peptide fragment covering this region was isolated by reversed-phase HPLC after digesting congerin I with endopeptidase Asp-N (1.7% enzyme by weight) (Takara Shuzo Co.) for 18 h at pH 8.5 and 37°C, and then sequenced by a protein sequencer (PSQ-1, Shimazu, Kyoto) and by the matrix assisted laser desorption ionization time of flight mass spectrometry (MALDI-TOF-MS).

Northern blotting and RT-PCR. Total RNAs were extracted from various tissues of conger eel and analyzed by northern blotting. In brief, glyoxal-denatured RNA samples (10 µg/lane) were electrophoresed through a 1.1% agarose gel, transferred to Hybond N+ (Amerham-Pharmacia), and hybridized with a digoxigenin-labeled congerin II antisense RNA probe prepared by in vitro transcription with a DIG RNA Labeling kit (T7 RNA polymerase) and subcloned plasmid for congerin II. After washing with 2× SSC/0.1% SDS at room temperature and twice with 0.1× SSC/0.1% SDS at 65°C, the signals were detected by using a chemiluminescent substrate, CDP-Star (Boehringer Mannheim) with a lumino image analyzer LAS-1000 (Fuji film). RT-PCR was done using Superscript one-step RT-PCR system (Life Technologies) with primers, Con I-5’ = 5’-ACCAATGCCCCATCATGAGTGAG-3’ and Con I-3’ = 5’-AAATATGCTGACGAACTG-3’, and Con II-5’ = 5’-ACCTGCTGCCATCATGAGT-3’ and Con II-3’ = 5’-AAATATGCTGACGCACTTTCT-3’ for congerins I and II, respectively.

Results

For screening of the cDNA library, PCR techniques were used to prepare the specific probes for congerins I and II (Fig. 1). The PCR resulted in the amplification of a 287-bp fragment for congerin I and a 332-bp fragment for congerin II. After subcloning the amplified DNAs into pBS II SK +, their nucleotide sequences were analyzed. The first screening of 1 × 10^6 clones encoding the

Fig. 1. PCR Primers for Congerin I and Congerin II.

(A) Amino acid sequences of congerins I and II. (B) Structures of the oligonucleotides used as PCR primers and sequence primers.
cDNA library with high stringency (0.1 × SSC, 0.1% SDS, 65°C) yielded 16 clones for conger I and 13 clones for conger II. The clone named Con I-10 contained a 5′-UTR of 21 base pairs (bp) followed by a single open reading frame of 411 bp, which codes for conger I, and a 3′-UTR of 213 bp. The nucleotide sequence of clone Con II-2 encoded conger II, and contained a 5′UTR of 48 bp, a coding region of 408 bp, and a 3′UTR of 210 bp (Fig. 2). No signal sequence was identified in the cDNAs of either conger I or conger II. The amino acid sequence predicted from the nucleotide sequence of the conger II cDNA was completely in agreement with that of the protein analyzed previously. However, the predicted amino acid sequence from conger I cDNA differed from the published sequence by one residue, the second from the C-terminus (indicated by underline in Fig. 2). Thus, the protein sequence was re-investigated by digesting conger I with endoprotease Asp-N, and by sequencing the peptide corresponding to the C-terminus which covers the region in question. Figure 3 shows the HPLC profile of the endoprotease Asp-N digest. The peptide N1 indicated by an arrow in Fig. 3 gave the amino acid sequence Leu-Thr-Leu-Val-Lys-Leu-Glu, which corresponds to the sequence from the 131st to 136th positions, and was identical to the sequence predicted from the cDNA sequence (Fig. 2). Further, this result was confirmed by MALDI-TOF MS. The molecular mass of peptide N1 (814.8 Da) determined by the MALDI-TOF-MS spectrum was in good agreement with the calculated value of 815.02. The molecular mass of conger I was 15,359 Da from the MALDI-TOF-MS, which was consistent with the calculated value, 15,359.23 Da (data not shown).

The nucleotide sequences of cDNAs for conger I

![Absorbance at 210 nm](image)

![Time [min]](image)
and congerin II are compared in Fig. 2. The nucleotide substitutions occur predominantly in the protein-coding region, indicating that the protein coding region is much more variable than the 5' and 3'-UTRs. The sectional identities between congerins I and II were 86.4% for the 5'-UTR, 73% for the protein-coding region and 88% for the 3'-UTR. To obtain further information about their molecular evolution of congerins I and II, the number of nucleotide substitutions per site (Ks), the number of nucleotide substitutions per synonymous site (Ks), and the number of nucleotide substitution per non-synonymous site (Ka) were calculated for the two cDNAs. Based on the neutral theory of molecular evolution, the Ka value never exceeded the Ks value in the average evolution. If Ka exceeded Ks (Ka/Ks > 1.0), it could be considered that substitution is accelerated in the protein-coding region. As shown in Table 1, the Ka/Ks values are 0.715 for the 5'-UTR, and 0.743 for the 3'-UTR. On the other hand, the Ka/Ks value for the protein-coding region (2.57) was greater than 1.0, indicating that nucleotide substitutions that caused amino acid changes occurred more frequently than did nucleotide substitutions that caused no amino acid changes.

The tissue distribution of mRNAs for congerins was examined by using the congerin cDNAs as probes. Northern blot analysis for congerin II showed positive signal not only in the skin tissues but also in the stomach tissues (Fig. 4A). To confirm the expression of congerins in the stomach tissues, RT-PCR was made for both skin and stomach tissues RNA using two pairs of primers (Con I-5' / Con I-3' for congerin I and Con II-5' / Con II-3' for congerin II). The expected DNA fragments, 450 bp for congerin I and 430 bp for congerin II, were amplified in both skin and stomach tissues (Fig. 4B). After subcloning these amplified DNAs, their nucleotide sequences were analyzed (data not shown). It was found that the amplified DNAs from stomach encoded congerins I and II, respectively. Thus, congerins I and II are considered to be expressed in the skin and stomach tissues. The kidney also showed distinct positive signals at a higher molecular size than congerin II (Fig. 4A), suggesting that the kidney may express some protein similar to congerins, may be including other galectins.

Discussion

The cDNAs encoding conger eel galectins named congerin I and congerin II were isolated and their nucleotide sequences were analyzed. The most interesting feature found in this study is that the protein coding regions in these cDNAs are much more variable than the untranslated regions. The percentage of the nucleotide sequence identities in the 5'-UTRs, the coding regions, and the 3'-UTRs are 86.4%, 73%, and 88%, respectively. These unusual structural features of congerin cDNAs seem to reflect an evolutionary phenomenon that has previously been seen in crotaline snake venom isozyme cDNAs and genes, but the phenomenon is the opposite to that of general isoprotein genes in which the protein-coding region are much more conserved than the UTRs.

Mathematical analysis between congerin I and congerin II cDNAs showed that the Ka/Ks value for the protein-coding region (2.57) is greater than 1.0, and the Ka/Ks value for the 5' and 3'-UTRs is 0.74. These observations suggest that non-synonymous substitution has occurred frequently in the protein-coding region, and
that congerins I and II have evolved via accelerated substitutions from their common ancestor gene. Why does the skin mucus of conger eel contain multiple galectins that are generated by accelerated evolution? The skin of fish is covered with a mucus layer, which faces the external surroundings and defends against infections from bacteria and parasites. Many components such as toxin, hemolysin, and lysozyme contained in the skin mucus of fishes are functional in the biodefense system. Galectins are also considered to play an important role, either by functioning independently or by cooperating with other components. We previously demonstrated that the conger can agglutinate the marine bacterium Vibrio anguillarum. Furthermore, congerins I and II showed different binding specificities for lactose derivatives, and showed different properties such as pH stability and thermostability. It is likely that the presence of multiple galectins in the skin mucus of fishes provides some advantages for the biodefense system. For example, with different forms and thus different carbohydrate specificities, galectins could respond to many different types of bacteria.

Congerins I and II share some structural characteristics such as an acetylated N-terminus and absence of cysteine residues and sugar chains. In this study, no signal sequence was observed in the amino acid sequences deduced from the cDNA sequences. However, congerins I and II are secretory proteins that are presumed to be produced by club cells and stored in the skin mucus, as is the case with eel lectins. The absence of a signal sequence suggests that congerins are secreted by a mechanism different from that of general secretory proteins with signal peptides. Signal sequences were also reported to be absent in mammalian and chicken galectin cDNAs. Furthermore, it has been reported that galectin is actively secreted from cells by direct translocation across the plasma membrane. The process of externalization of galectin-1 from muscle cells during in vivo development has been well studied by immunohistochemical techniques. Furthermore, other galectins are also secreted to external cells. For example, a 14-kDa of chicken galectin has been found in internal epithelial cells and directly shown to be secreted into the intestinal lumen. A galectin from Xenopus skin has been found to be secreted by a specialized holocrine mechanism. Mammalian galectins are known to be widely distributed in mammalian tissues. It was reported that galectin-1 was abundant in skeletal, cardiac muscle, motor and sensory neurons, fibroblasts, kidney, and placenta. In this study, we demonstrated that congerin II is expressed in the gastric tissues in addition with the skin including epidermal tissues such as mucus cells, saciform cells, and club cells.

In the galectin family, the CRD including His-44, Arg-48, His-52, Asp-54, Asn-61, Trp-68, Glu-71, and Arg-73 (the numbers are based on the sequence of galectin-1 from bovine spleen) are highly conserved, and these residues are involved in the molecules’ sugar binding activities. From the X-ray structure of the galectin-1-N-acetyl lactosamine complex, the amino acid residues Glu-73 and Arg-75 interact with the C-3 hydroxyl groups of N-acetyl lactosamine in the CRD of bovine galectin-1. His-44 and Arg-48 make important hydrogen bond interactions with the galactose 4-OH, and Asn-61 and Glu-71 interact with galactose 6-OH. Asn-46 in the mammalian galectin-1 has been found to be involved in the interaction with the 3-OH through a water molecule and the 4-OH of lactose. Trp-68 is located in the bottom of the CRD cleft to make a stacking interaction with the galactose moiety. For congerins I and II, seven amino acids (His-44, Arg-48, Asp-54, Asn-61, Trp-68, Glu-71 and Arg-73) are well conserved, although two amino acid residues at positions 46 (Asp instead of Asn-46) and 52 (Gly instead of His-52) are substituted. These conserved amino acids in the CRD of congerins I and II are probably responsible for their binding specificity for β-galactoside. On the other hand, congerins I and II have variable amino acid residues generated by accelerated evolution, which are probably related to their different binding properties with respect to C-3-modified methyl lactose derivatives and glycoprotein.

Multiple prototype galectins have been found in mammals and chickens, and their nucleotide sequences have been analyzed. Although the human isogalactosins (L-14-I, L-14-II, and HKL14) and chicken isogalactosins (C-14 and C-16) showed only 40 to 50% identities in their amino acid sequences, their nucleotide sequences show no significant evolutionary phenomenon similar to that seen in congerin cDNAs. The sectional nucleotide identities between C-14 and C-16 were almost the same, that is, 63% for the 5’UTR and 61% for the coding region, although the 3’ nucleotide sequence of C-16 was lacking. These results suggest that conger eel galectins differ from other galectins in having diversified their amino acid sequences in an accelerated manner. Similar evolutionary behavior, i.e., accelerated substitution, has previously been observed in the cDNAs and genes encoding snake venom isozymes and isotoxins. Thus, common evolutionary mechanisms for accelerated substitution may be present in lower vertebrates, snake toxins, and conger eel galectins.

References


