THE JOURNAL OF BIOLOGICAL CHEMISTRY
© 2001 by The American Society for Biochemistry and Molecular Biology, Inc.

Vol. 276, No. 23, Issue of June 8, pp. 20624–20632, 2001
Printed in U.S.A.

Molecular Patterning of the Oikoplastic Epithelium of the Larvacean Tunicate *Oikopleura dioica**

Received for publication, January 17, 2001, and in revised form, February 28, 2001
Published, JBC Papers in Press, March 6, 2001, DOI 10.1074/jbc.M100438200

Fabio Spada‡, Hanno Steen§, Christofer Troedsson‡, Torben Kallesøe‡, Endy Spriet‡, Matthias Mann§, and Eric M. Thompson‡¶

From the ‡Sars International Centre for Marine Molecular Biology, Bergen High Technology Centre, N-5008 Bergen, Norway and the §Protein Interaction Laboratory, Department of Biochemistry and Molecular Biology, University of Southern Denmark-Odense, DK-5230 Odense M, Denmark

Appendicularia are protochordates that rely on a complex mucous secretion, the house, to filter food particles from seawater. A monolayer of cells covering the trunk of the animal, the oikoplastic epithelium, secretes the house. This epithelium contains a fixed number of cells arranged in characteristic patterns with distinct sizes and nuclear morphologies. Certain house structures appear to be spatially related to defined, underlying groups of cells in the epithelium. We show that the house is composed of at least 20 polypeptides, a number of which are highly glycosylated, with glycosidase treatments resulting in molecular mass shifts exceeding 100 kDa. Nanoelectrospray tandem mass spectrometric microsequencing of house polypeptides was used to design oligonucleotides to screen an adult *Oikopleura dioica* cDNA library. This resulted in the isolation of cDNAs coding for three different proteins, oikosin 1, oikosin 2, and oikosin 3. The latter two are novel proteins unrelated to any known data base entries. Oikosin 1 has 13 repeats of a Cys domain, previously identified as a subunit of repeating sequences in some vertebrate mucins. We also find one repeat of this Cys domain in human cartilage intermediate layer protein but find no evidence of this domain in any invertebrate species, including those for which entire genomes have been sequenced. The three oikosins show distinct and complementary expression patterns restricted to the oikoplastic epithelium. This easily accessible epithelium, with differential gene expression patterns in readily identifiable groups of cells with distinctive nuclear morphologies, is a highly attractive model system for molecular studies of pattern formation.

Appendicularia, or larvaceans, are pelagic tunicates (Urochordata) that feed on dissolved organic carbon and microorganisms by filtering seawater through a transparent mucous structure called the house. The name "larvaceans" derives from the fact that the adult animal resembles the larval form more closely than in the other two classes of Urochordata. In the latter two classes, metamorphosis drastically redesigns the body plan, whereas in appendicularia, it consists of a simple switch in tail position (tailshift) from a straight posteriorly directed orientation to a definitive arrangement, where the tail is orthogonal to the trunk and retains the notochord as its axial structure. For this reason, appendicularia are thought to have diverged earlier from the chordate ancestor than their sister classes, and recent phylogenetic analyses of 28 and 18 S rRNA genes confirm this view (1, 2). Appendicularia have three features common to all chordates at some stage in the life cycle: gill slits, a tubular nerve chord, and a rod-shaped notochord. The appendicularia and vertebrates form a robust monophyletic unit, with two *Oikopleura* species forming a sister group to all vertebrates (2).

The term appendicularia refers to the appendices of the animal, namely their houses. Appendicularian houses are sophisticated filtration devices with a very complex architecture, featuring different compartments, valves, septa, and two sets of filters (3–5). The house is secreted as a rudiment by the oikoplastic epithelium, a specialized single-layered organ that covers the trunk of the animal (see Fig. 1). The house rudiment is expanded by means of specific movements of the tail, which is eventually withdrawn inside the house, such that the entire animal is contained within this structure. Sinusoidal movements of the tail control the flow of water through the house filters. The cells of the oikoplastic epithelium stop dividing before secretion of the first house. Their number is fixed for a given species of the genus *Oikopleura*, being about 2,000 in *Oikopleura dioica* (6–8). Frequent replacement of the house is needed to maintain sufficient filtration rates (9), and the cells of the oikoplastic epithelium sustain high rates of house protein synthesis and secretion from shortly before filter feeding begins until the end of the life cycle. For *O. dioica*, a new house is synthesized every 3–4 h over a life cycle of 5–6 days at 15 °C. During the life cycle, which is extremely short for a complex metazoan, the animal grows about 10-fold in size, and there is a concomitant increase in the size of the oikoplastic cells. Cell growth is paralleled by an increase in the sizes of the nuclei and in their DNA content (10).

It is not known whether the increase of DNA content represents partial or total amplification of the genome. During this process, however, the shape and size of the terminally differentiated cells of the oikoplastic epithelium becomes highly variable. These vary from round and relatively small to lobate, oblong shapes attaining a large size. The complexity of the house structure is mirrored by the diverse patterns of the oikoplastic cells. Thus far, different fields of cells have been identified within the epithelium on the basis of nuclear morphology, size, and pattern formation (6–8, 10, 11). A correlation between these morphologically identified fields of cells and

* This work was supported by the Norwegian Research Council and Ministry of Education. Work in the Protein Interaction Laboratory was funded by a grant from the Danish National Research Foundation to the Center for Experimental BioInformatics. The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

The nucleotide sequence(s) reported in this paper has been submitted to the GenBank™/EBI Data Bank with accession number(s) AJ308491, AJ308492, AJ308493, AJ308494, and AJ308495.

¶ To whom correspondence should be addressed. Tel.: 47-55-58-43-46; Fax: 47-55-58-43-05; E-mail: Eric.Thompson@sars.uib.no.

the production of specific structures of the house seems apparent for the two sets of filters, the inlet filter being secreted by the field of Eisen and the food concentrating filters being secreted by the field of Fol (see Fig. 1). These functional correlations remain crude when compared with the complexity of the filters and the patterns of cells held responsible for their secretion. The structural characteristics of different parts of houses of the genus *Oikopleura* have been studied by electron microscopy (4), but very little information, based primarily on histochemical staining, is available concerning their biochemical composition and properties (5).

Thus, *O. dioica* has a monolayer of accessible epithelium, with low and constant cell number, that serves as a template for the synthesis of diverse house structures from clearly defined regions of cells with distinct nuclear morphologies. This is an attractive model system for investigating: (i) coordinate regulation of gene expression, (ii) cell-cell interactions involved in pattern formation, (iii) gene/genome amplification, and (iv) the role of nuclear architecture in regulating gene expression. However, for the oikoplastic epithelium of *Oikopleura* to become a *bona fide* model for studying the questions posed above, it remains to be established that the structures in the house are synthesized from a diverse array of gene products with expression patterns restricted to different regions of cells. Alternatively, relatively few genes could be involved, with their products undergoing differential post-translational modifications.

In the present study we report partial biochemical characterization of the house polypeptides of *O. dioica* and the cloning of cDNAs encoding four of them. We demonstrate that the genes coding for these cDNAs are expressed in well defined complementary regions of the oikoplastic epithelium. One of the cDNAs codes for a very large polypeptide, showing distant homology to vertebrate gel-forming mucins and human cartilage intermediate layer protein, and the corresponding gene is expressed only in the 14 giant cells of the Fol region.

EXPERIMENTAL PROCEDURES

Appendicularia Culture—Animals were collected from fjords around Bergen, Norway and were cultured in 6-liter beakers with constant stirring at 14–15 °C. The duration of the life cycle was 5–6 days. Mature males and females were placed together in 4-liter volumes of seawater and allowed to spawn. Metamorphosis took place 12–14 h after spawning, at which time the animals expanded their first house and began filter feeding. The animals then grew continuously and on day 5 the gonads increased dramatically in size, surpassing the size of the trunk of the animal at spawning. Cultures were diluted 1:6 on day 1 and 1:1 on day 2, and then animals were transferred to clean sea water by pipetting on each of the following days. The appendicularia were maintained on cultured algal strains of *Isochrysis galbana*, *Chaetoceros calcitrans*, and *Synecococcus* sp., supplemented with *Rhodomonas* sp. and/or *Tetraselmis* sp. from day 3 onward.

House Protein Solubilization and Deglycosylation—To collect house rudiments free from contaminating particles and microorganisms, adult animals (day 4–5) were forced out of their houses and transferred to filtered seawater. House rudiments were then removed as soon as the animals started expanding them. Aliquots of 50–200 house rudiments were stored at −80 °C. After thawing, aliquots were centrifuged at 12,000 × *g* for 10 min at 4 °C to remove excess sea water. Rudiments were then fragmented by making a few holes with a 27-gauge needle in the bottom of the tube and centrifuging this into a second collecting tube. House rudiments were then heated at 100 °C for 5 min in 1% SDS. The denatured rudiments were diluted 5-fold with 25 mM sodium phosphate buffer, pH 7.2, 12.5 mM EDTA, 0.625% CHAPS,[1] and 1.25% β-mercaptoethanol and heated again at 100 °C for 5 min. After cooling, protease inhibitors (a leupeptin, pepstatin, E-64, bestatin, (4-[2-aminoethyl]benzenesulfonylfluoride, and aprotinin mixture from Sigma diluted 1:100) were added, samples were centrifuged at 12,000 × *g* for 10 min at room temperature, and the supernatant was recovered. Glycosidases were added to a supernatant obtained from 50 house rudiments as follows: 0.2 unit of peptide-*N*-glycosidase F (Roche Molecular Biochemicals), 5–20 milliunits (as indicated) of α2–3,6,8-neuraminidase (Calbiochem), and 2 milliunits of bovine serum albumin-free *O*-glycosidase (Roche Molecular Biochemicals). When larger aliquots were processed, enzyme amounts were scaled up proportionally. Incubation was at 37 °C for at least 18 h. Undigested samples were incubated under the same conditions.

Protein Electrophoresis—For one-dimensional SDS-polyacrylamide gel electrophoresis (PAGE) samples were loaded directly, or in the case of large solubilization volumes, precipitated with 20% trichloroacetic acid, and then redissolved in loading buffer. Gel and buffer compositions were according to Shägger and von Jagow (12), except that 16-cm-long gels with a 5–16% polyacrylamide gradient (3% crosslinking) were used to maximize resolution over a wide molecular mass range. Gels were stained with minor modifications of the combined alcian blue-periodic acid oxidation-silver staining procedure of Møller and Poulsen (13) or by neutral silver staining (14).

Two-dimensional electrophoresis was performed with isoelectrofocusing on immobilized pH gradients as the first dimension and SDS-PAGE as the second dimension. Deglycosylated samples were precipitated with trichloroacetic acid as above and resuspended in 0.5% immobilized pH gradient buffer, pH 3–10 (APBiotech), 2% CHAPS, 7 M urea, 2 M thiourea, 5 mM tributyl-phosphine, and 0.005% bromphenol blue. Dry immobilized pH gradients gel strips pH 3–10 NL (nonlinear pH gradient; 7 cm long; APBiotech) were rehydrated with resuspended samples overnight, and first dimension electrophoresis was performed with a Multiphor II apparatus (APBiotech). Electrofocused gel strips were then equilibrated in second dimension loading buffer (50 mM Tris-HCl, pH 6.8, 6 M urea, 2% SDS, 30% glycerol, 0.05% bromphenol blue) and applied on a vertical second dimension gel with a 5–16% polyacrylamide gradient as described for one-dimensional electrophoresis, except that the resolving gel was 11 cm long and no stacking gel was used. Two-dimensional gels were stained with the same neutral silver staining procedure as described for one-dimensional gels. Spots (two-dimensional) and bands (one-dimensional) corresponding to the added glycosidases were identified by comparison with gels run with enzyme blanks under identical conditions.

Peptide Sequencing—Protein bands were excised from silver-stained one-dimensional polyacrylamide gels and processed as described (14). After reduction and alkylation, overnight in-gel digestion was carried out at 37 °C with an excess of sequencing grade trypsin (Roche Molecular Biochemicals). After digestion, the supernatant was acidified with formic acid, loaded onto a Poros R2 (Perseptive Biosystems, Framingham, MA) micro-column, and desalted (15). The peptides were then eluted with 60% methanol, 5% formic acid directly into a nanoelectrospray needle (MDS Protana, Odense, Denmark). Peptides were sequenced by nanoelectrospray tandem mass spectrometry performed on a prototype quadrupole time-of-flight mass spectrometer (AB/MDS Sciex, Toronto, Canada) (16) equipped with a nanoelectrospray source (MDS Protana).

Library Construction and Screening—Total RNA from day 5 animals was extracted with the TRIzol Reagent (Life Technologies, Inc.), and the poly(A)$^+$ fraction was isolated from ~200 μg of total RNA by two rounds of oligo(dT)-cellulose chromatography with the Poly(A)$^+$ Quik mRNA isolation kit (Life Technologies, Inc.). This poly(A)$^+$ RNA was the template for cDNA synthesis using the ZAP Express cDNA synthesis kit, and the cDNA was cloned and packaged using the ZAP Express cDNA Gigapack III Gold Cloning Kit (Stratagene) except that cDNA size selection was performed with SizeSep 400 Spun Columns (APBiotech). A primary library of 3×10^5 plaque-forming units was obtained and amplified once. A plating of 3×10^5 plaque-forming units of the amplified library was screened with degenerate oligonucleotides corresponding to the peptide sequences shown in *bold type* in Fig. 3. The oligonucleotides were synthesized at Interactiva Biotechnologie (GmbH): G2, 5′-TGGTAYGGIGAYGAYCARGG-3′; G10, 5′-CAYTTYCCITGGTAYAAYAA-3′; H5, 5′-GAYACIGCITAYCAYGGITT-3′; I1, 5′-GARTTYTGYGGIGTIGAYTT-3′; and N6, 5′-TAYGAYAAYGAYCARGCIGC-3′, where Y = C or T, R = G or A, and I = inosine. Oligonucleotide nomenclature is based on the labeling of the protein bands from which they were derived. Oligonucleotides were end-labeled with [γ-^{32}P]ATP using T4 polynuclotide kinase (New England Biolabs) and pooled in hybridization buffer in relative concentrations such that each possible sequence of any of the degenerate sequences was present at a concentration of 31,25 pM. Filters were prehybridized in 6× SSC, 10 mM

[1] The abbreviations used are: CHAPS, 3-[(3-cholamidopropyl)dimethylammonio]-1-propanesulfonic acid; PAGE, polyacrylamide gel electrophoresis; MOPS, 4-morpholinepropanesulfonic acid; PNGase F, peptide-*N*-glycosidase F; GalNAc, N-acetylgalactosamine; bp, base pair(s); CILP, cartilage intermediate layer protein.

sodium phosphate, 5× Denhardt's solution, 0.5% SDS, and 100 μg/ml sheared herring sperm DNA and hybridized in 6× SSC, 10 mM sodium phosphate, pH 7.2, 1× Denhardt's solution, 0.5% SDS, and 100 μg/ml yeast tRNA at 37 °C for 48 h. Washing was at 37 °C in 6× SSC.

Polymerase chain reaction screening was performed with the standard T3 primer as the vector-anchored primer and with either primer A5′ (5′-CAACGCTGAAGTCATCATC-3′), G5′ (5′-AAGCCTTGCATGCGTGC-3′), or H5′ (5′-GGATTTGAACCACGACAAC-3′) for amplifying sequences upstream of clones A1, G1, and H, respectively. A first set of primer extention reactions (linear amplification) were carried out in a total volume of 50 μl with 5 μl of amplified library as template, 0.2 μM specific primer, 200 μM of each dNTP, and 1 unit of DyNAzyme II DNA polymerase (Finnzyme) in DyNAzyme reaction buffer (10 mM Tris-HCl, pH 8.8, at 25 °C, 50 mM KCl, 1.5 mM $MgCl_2$, and 0.1% Triton X-100). After 35 cycles, 5 μl of these reactions were used as templates for exponential amplification reactions under the same conditions as above, but with the inclusion of 1 μM of T3 primer. Amplified fragments were digested with the appropriate restriction enzymes, gel purified, and cloned in pBluescript II.

Northern Blotting and Whole Mount in Situ Hybridization—Total RNA was prepared from day 4–5 animals by acid guanidinium thiocyanate-phenol-chloroform extraction, and the poly(A)$^+$ fraction was isolated by two rounds of oligo(dT)-cellulose chromatography. Aliquots of 500 ng of the poly(A)$^+$ fraction were electrophoresed through a 2.2 M formaldehyde, 1% agarose gel and transferred to a positively charged nylon membrane (Hybond-N+, APBiotech) by capillary blotting. The RNA was fixed to the wet membrane by UV cross-linking with 150.000 μJ/cm^2 at 254 nm with a Hoefer UVC 500 Ultraviolet Cross-linker (APBiotech). Purified DNA fragments were radioactively labeled by random priming with the T7 Quick Prime kit (APBiotech), and unincorporated nucleotides were removed by Sephadex G-50 spun column chromatography (APBiotech). Membranes were prehybridized and hybridized in 250 mM sodium phosphate buffer, pH 7.2, 7% SDS, 1 mM EDTA, 1% bovine serum albumin and washed with 20 mM sodium phosphate buffer, pH 7.2, 1% SDS, 1 mM EDTA. Hybridized blots were analyzed with a FLA 2000 phosphoimager and Image Gauge software (Fuji).

For whole mount *in situ* hybridization, sense and antisense RNA probes were synthesized by *in vitro* transcription of linearized plasmids containing the appropriate inserts with either T7 or T3 RNA polymerase (Promega) in the presence of digoxigenin-labeled UTP (digoxigenin RNA Labeling Mix; Roche Molecular Biochemicals). Day 2–3 animals were fixed with 4% paraformaldehyde in 100 mM MOPS, pH 7.5, and 500 mM NaCl for 20 min and digested with 10 μg/ml proteinase K in the same buffer for 3 min at 37 °C. Specimens were post-fixed for 5 min with the same fixing solution as above. Prehybridization (2 h) and hybridization were carried out in 50% deionized formamide, 5× SSC, 2% blocking reagent (Roche Molecular Biochemicals), 0.1% Triton X-100, and 9.2 mM citric acid at 60 °C. Probes were diluted at 400 ng/ml in hybridization mixture and denatured at 85 °C for 5 min before being added to the specimens and incubated overnight. Samples were washed once with 4× SSC, 50% formamide, 0.1% Triton X-100 for 30 min at 37 °C; once with 2× SSC, 2 mg/ml bovine serum albumin, 0.1% Triton X-100 for 30 min at 37 °C; and twice with 0.5× SSC, 2 mg/ml bovine serum albumin, 0.1% Triton X-100 for 30 min at 65 °C. Hybridized specimens were blocked with 1% blocking reagent in phosphate-buffered saline with 0.1% Triton X-100 for 1 h at room temperature and incubated in the same medium containing 2 μg/ml of either fluorescein- or rhodamine-conjugated anti-digoxigenin Fab fragments (Roche Molecular Biochemicals) for 40 min at room temperature. After washing three times with phosphate-buffered saline with 0.1% Triton X-100 for 30 min, specimens were mounted in Vectashield mounting medium (Vector Laboratories) and analyzed with a Leica TCS laser scanning confocal microscope equipped with a 40× Leica oil immersion objective (numerical aperture 1.25–0.75). Images were acquired with Leica PowerScan software. Parallel hybridizations with sense RNA probes were run as controls.

RESULTS

Partial Characterization of House Proteins—To avoid contamination of house protein preparations by algae and bacteria, all preparations were made using noninflated prehouse rudiments (Fig. 1). Several solubilization procedures were tested, including sonication, detergent extraction, and guanidine HCl extraction. Denaturation by heating in the presence of 1% SDS, followed by titration of SDS with CHAPS was the most effective. One-dimensional SDS-PAGE of solubilized and reduced house material yielded several bands within a range of molecular mass from 5 to over 200 kDa (Fig. 2). Among these,

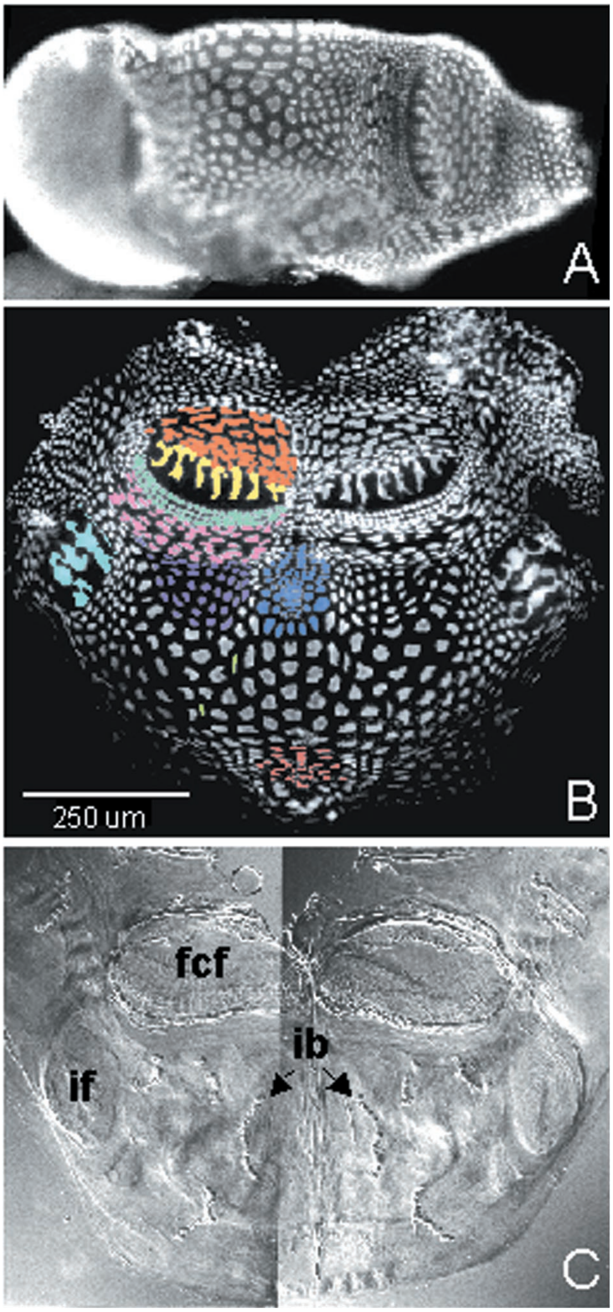

FIG. 1. **Oikoplastic epithelium and house rudiment of *O. dioica*.** *A*, trunk of a fixed animal showing nuclei stained with Hoechst 33258. The mouth is to the *right*, with the maturing gonad to the *left*. The gonad is also stained because it contains maturing sperm. *B*, epithelial spread stained with Hoechst 33258. The spread was obtained by removing the tail, the gonad, and the internal organs from a fixed animal and cutting the oikoplastic epithelium longitudinally along the mid-ventral line. The mouth is at the *top*, and the posterior part of the epithelium at the *bottom*. Different colors indicate regions of the epithelium that have been previously defined on the basis of distinct nuclear morphologies and spatial links with house structures. The field of Fol includes the anterior Fol (*orange*), the seven giant cells (*yellow*), the Nasse cells (*turquoise green*), and the posterior Fol (*pink*). The field of Eisen includes the chain of pearls and the seven giant cells (*light blue*). The field of Martini (*violet*), the anterior (*dark blue*) and posterior rosettes (*red*), and the Leuckart cells (*green*) are also indicated. *C*, house rudiment spread. The house rudiment was isolated from the animal and cut the same way as the epithelium in *B*. *fcf*, food concentrating filter; *if*, inlet filter; *ib*, bioluminescent inclusion bodies.

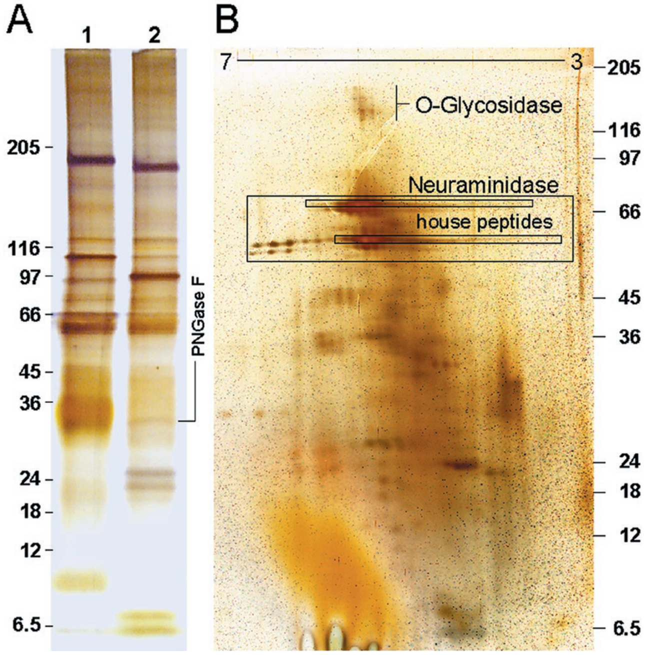

FIG. 2. **Solubilized, denatured, and reduced house polypeptides were digested with different combinations of glycosidases and electrophoretically resolved.** *A*, one-dimensional SDS-PAGE of house polypeptides without (*lane 1*) and with PNGase F treatment (*lane 2*). In each lane, material corresponding to 50 house rudiments was loaded, and the bands were revealed with combined periodic acid oxidation-alcian blue-silver staining. Following PNGase F treatment, there was a significant decrease in molecular mass of several bands, including the disappearance of a broad, prominent band at 36 kDa and the appearance of lower molecular mass bands. The positions of molecular mass standards and the PNGase F band are indicated. *B*, two-dimensional electrophoresis of house polypeptides simultaneously digested with PNGase F, neuraminidase, and *O*-glycosidase. The gel was stained with neutral silver. Spots corresponding to added enzymes are indicated and were identified by running a parallel two-dimensional gel with enzymes only. PNGase F spots migrate in a basic pH range outside the part of the gel shown here. The two *rectangular boxes* inside the neuraminidase box mark house polypeptides with a wide range of pI heterogeneity. Note the disappearance of several bands with molecular masses over 70 kDa compared with PNGase F treatment alone in *lane 2* of *A*. The nonlinear pH gradient is indicated on *top* of the two-dimensional gel, and the positions of molecular mass markers are on the *right side*.

a broad, prominent band at 36 kDa and several sharper bands appeared as clear bands upon neutral and particularly upon ammoniacal silver staining (data not shown). Both molecular mass heterogeneity and negative staining are associated with highly glycosylated proteins, such as proteoglycans and mucins (13, 17). Following combined periodic acid oxidation-alcian blue-silver staining, specific for highly glycosylated proteins, most of these bands were positively stained (Fig. 2). Different glycosidases were employed to reduce polypeptide heterogeneity. After digestion with peptide-*N*-glycosidase F (PNGase F), which hydrolyzes the linkage between asparagine and *N*-acetylglucosamine common to all *N*-linked oligosaccharides, a shift to lower molecular masses was observed for several polypeptides (Fig. 2*A*). The heterogeneous band at 36 kDa disappeared, probably giving rise to some of the lower molecular mass species that appeared following this treatment. After simultaneous digestion with PNGase F and a broad spectrum neuraminidase, all bands larger than 70 kDa disappeared, but neuraminidase action produced a heavy background on gels stained with the combined alcian blue-periodic acid oxidation-silver procedure. When the gel was neutral silver stained, the pattern was very simple, with only a few peptide bands detectable. Two dimensional electrophoresis of a sample treated the same way showed that at least 20 polypeptide species were detected and that pI heterogeneity was still present in some of them (Fig. 2*B*). The electrophoretic patterns from repeated independent preparations were consistent, and no differences were observed when these preparations where digested in the presence or absence of a broad spectrum mixture of protease inhibitors. Thus, these deglycosylation treatments resulted in substantial decreases in molecular mass profiles, indicating extensive glycosylation of house proteins, and the large shifts resulting from treatment with neuraminidase alone suggest that sialic acid moieties were abundant.

In contrast to PNGase F and neuraminidase treatments, addition of *O*-glycosidase (endo-α-acetylgalactosaminidase) to house rudiment preparations had no detectable effect on the electrophoretic pattern. *O*-Glycosidase hydrolyzes the bond between serine or threonine and acetylgalactosamine (GalNAc) in the disaccharide core unit Galβ1–3GalNAc when the latter is not substituted by sialic acid or other oligosaccharides. Therefore, our result is not surprising, because the Galβ1–3GalNAc core unit is only present in a subset of *O*-linked oligosaccharides and a significant amount of this linkage unit would be required to see detectable molecular mass shifts. Most of the glycosaminoglycans found in proteoglycans and the oligosaccharide chains of mucins are linked to serine and threonine through different saccharide moieties.

cDNA Cloning and Sequencing—To isolate cDNA clones encoding house components, bands prepared from one-dimensional SDS-PAGE separation of PNGase F digested house proteins were subjected to nanoelectrospray-tandem mass spectrometry. Neuraminidase treatment was omitted because it added several enzyme bands/spots that co-migrated with house peptide bands/spots in both one- and two-dimensional separations. The most abundant polypeptide bands were cut out of the gel, digested with trypsin, eluted, and microsequenced. Two bands at 96 kDa were insufficiently separated and were processed as a single band labeled *G* in Fig. 3. The peptide sequences obtained by microsequencing are shown in Fig. 3.

One of the peptide sequences from band A (*bold type* in Fig. 3) matched perfectly with the partial sequence of a 3372-bp cDNA clone (clone A_1) from an *Oikopleura* cDNA shotgun sequencing effort.[2] To isolate additional clones, degenerate oligonucleotides were designed from selected peptide sequences obtained from bands G, H, N, and I (*bold type* in Fig. 3). These oligonucleotides were radioactively labeled, pooled, and used as probes to screen a cDNA library. Thirty-six clones were isolated, and partial sequencing revealed that the amino acid sequences deduced from reading frames present in 20 of them matched with microsequenced peptides. Eighteen of these clones encoded polypeptides matching peptides derived from band G (G clones), one for a polypeptide matching peptides from band H (clone H), and one matched all seven peptides from band A (clone A_2), although no probe sequence derived from band A was used in this screening (see below). The nucleotide sequence of this clone was identical to clone A_1 but had an additional 1173 bp at the 5′ end. The partial sequences of the remaining 14 clones did not show significant matches to any of the sequenced peptides. BLAST-X homology searches showed that seven of these clones encoded polypeptides similar to known mammalian proteins with functions not obviously related to house polypeptides (*E* values between 10^{-15} and 10^{-110} on nonredundant NCBI data bases). The remaining seven clones did not show significant homology to any known protein in the same data bases (*E* value $\leq 10^{-11}$). Because only partial sequences of the latter clones were ob-

[2] H.-C. Seo, unpublished data.

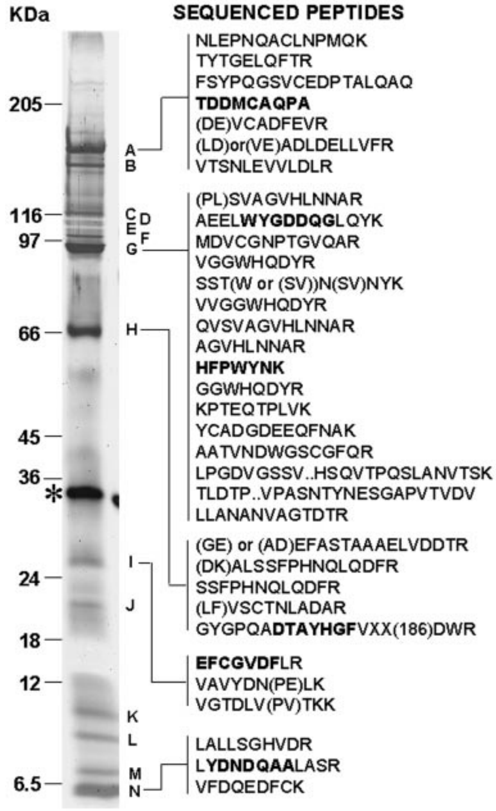

FIG. 3. **Neutral silver-stained one-dimensional SDS-PAGE of a 200-house rudiment sample digested with PNGase F alone, from which polypeptide bands were cut and microsequenced.** Peptide sequences derived from labeled bands are shown. Where two letters are in *parentheses*, the order of those residues could not be established. *XX* represents two residues that could not be identified, and the *number* in *parentheses* indicates the sum of their molecular masses. Stretches of peptide sequences from which oligonucleotide probes were derived are shown in *bold type*. The PNGase F band is marked by an *asterisk*, and the positions of molecular mass standards are indicated.

tained, it is possible that unsequenced portions match some of the sequenced house peptides. These clones have not yet been further characterized.

On the basis of alignments of partial sequences, restriction pattern analysis, and matches with the peptides derived from band G, there were three distinct types of G clones. The longest clones of each type, G1, G2a, and G2b were 2542, 2552, and 2559 bp long, respectively, and were completely sequenced. Clones G1 and G2 share an overall identity of 71% at the nucleotide level and code for polypeptides that show an identity of 70% (Fig. 4*A*). Clone G2a differs from G2b by only a few residues, a difference probably caused by allelic polymorphism. Out of the 18 G clones, 13 were of the G1 type, 3 of the G2a type, and 2 of the G2b type. This is consistent with the presence of multiple bands in the gel slice labeled G. Clone H, 2189 bp, was also completely sequenced. Translations of clones A_1, G1, G2a, G2b, and H showed uninterrupted reading frames starting from the 5′ end, but there were no putative starting methionine residues and no signal peptides predicted for any of them, indicating that all four clones were partial cDNAs.

To obtain full-length coding sequences for clones A_1, G1, and H, both primer extension and rescreening of the library with 5′ fragments of each clone were carried out. Full-length clones were obtained for both A and H. The 6996-bp nucleotide sequence of clone A contained an open reading frame that was 2301 amino acids long and started 10 nucleotides after the 5′ end of the cDNA insert. Two different prediction methods indicated the presence of a signal peptide with a cleavage site between amino acids 14 and 15 (Fig. 4*B*) (18, 19). Interestingly, the predicted molecular mass of the polypeptide coded by clone A is 256 kDa, whereas band A was only 180 kDa prior to and 170 kDa following PNGase F digestion (Fig. 2*A*). In addition, upon treatment with neuraminidase, the molecular mass of band A decreased by at least 100 kDa, because no polypeptide larger than 70 kDa was detected after neuraminidase digestion (Fig. 2*B*). Because glycosidase digestion patterns on extracellular house rudiments were the same, both in the presence and absence of protease inhibitors, it is unlikely that this difference in molecular mass can be accounted for by proteolytic degradation, suggesting that polypeptide A is post-translationally processed *in vivo*.

The full-length sequence for clone H was 2396 bp in length with an initiation codon at position 43. The predicted molecular mass of the polypeptide translated from this frame was 67 kDa, the same as that estimated by SDS-PAGE for band H. For clone G1, neither polymerase chain reaction extension nor the screening of 20 additional clones in the 5′ probing of the library yielded any sequence upstream of the 5′ end of the existing G1 clone. This suggests that secondary structure prevented reverse transcription through the 5′ region of the G1 transcript.

Nomenclature—Using the Greek *Oikos*, for house, as a root, we have assigned the following names to the clones presented in this study: clone A = oikosin 1, clone G1 = oikosin 2A, clone G2a = oikosin 2B1, clone G2b = oikosin 2B2, and clone H = oikosin 3. The EMBL data base accession numbers for the nucleotide sequences are AJ308491, AJ308492, AJ308493, AJ308494, and AJ308495, respectively.

Sequence Analysis—Oikosin 1 contained 13 motifs similar to the Cys subdomains present in three members of the 11p15 human mucin gene cluster (MUC2, MUC5B, and MUC5AC), in their rodent homologues, and in other related mammalian mucin cDNAs (HGM-1 and PGM-2A) (20–22). A similar motif is also present in human cartilage intermediate layer protein (CILP) (23) and two expressed sequence tag clones from *Xenopus laevis*. Several blocks of highly conserved residues were evident in the alignment of the 13 motifs of oikosin 1 and the cysteine subdomains found in these proteins (Fig. 5*A*). Among these, 6 of the 10 invariant cysteine residues and the putative mannosylation consensus sequence W*XX*(W/F) in the mucin Cys subdomains were present in all repeats of oikosin 1. The CILP Cys domain lacks the same 4 cysteine residues as the oikosin 1 repeats and, in addition, lacks the first and the penultimate cysteines as well. In the CILP Cys domain the second tryptophan in the putative mannosylation consensus is also substituted with leucine, a replacement known to reduce the efficiency of C-mannosylation of RNase 2 from 73% to 6% in HEK293 cells (24). Despite this, CILP is the closest match for 11 of 13 of the oikosin 1 repeats when they are used as queries in BLAST-P searches and when the PRODOM data base is searched with a profile obtained from all 13 repeats. CILP also differs from both the mammalian mucins and oikosin 1 in that it contains only one copy of the Cys domain. Fig. 5*B* shows schematic representations of some of the proteins whose Cys (sub)domains are aligned in Fig. 5*A*. These motifs were named Cys subdomains because they were first identified in the 11p15 human mucin cluster where they form a part of larger repeating units (20). No significant homologies outside the Cys (sub)domains were found among the proteins in Fig. 5, making the nomenclature "Cys subdomain" inappropriate for CILP and the oikosin 1 motifs. To our knowledge, Cys (sub)domain-like motifs had not been previously identified in CILP or in any known invertebrate proteins.

The function of the Cys (sub)domain is unknown. Mammalian mucins MUC2, MUC5B, and MUC5AC, are gel-forming

A

Clone A

```
MLLLSALLLGLAHGYSGTCRTIDPSRWTTGLTTFWGGAEQAQLTIDFDNDNKFQDFILNDETYVLVHFLVNVDTLTLNPTYPINQDCIQGTPTGQENT
NGCLDCVCSSCSCGYGFTEQKVLYMEVDDDESGRDVMSGQPSQYFEFQLLNKNVGFFIIDTVNFVEVCQGISPTVTGTTATNSPSDIAPVDDIVAELN
PGYAQGNVHSLATWRRTMVLHFAVEGATPVRENDYVLVSFDGEISILDFWSPFKEAVQVGDTVNENVWLFRFDADYNIAKSGSGGNWYGNMHLNFIGS
NWQRPAVQWVCVVEDPSNKGGATTTTTTTTPAPTTTTPTTTPAPPADPVKYPLSAGTCYSGNWTWWEQTDLVDDGIDNELYSNSIMRCQYPQIAAMKM
MDTEDSDLTETFQVINFDLDAGSGTCDNVDQDPLDLCHEYMVSFCCEDCCPILEITADPTNPPDLIEYYENYPGIYKLQTELFNDAITYRQVLGKDDS
GNDIFGEGIVYYWEDVGWLLGANTFTYSYYSDGIGERCPQFAPKNWTNRVYGTELNVECIPVVPSCNDVECVENAFCVMTGWGPECMCIPGYNLYEDG
SCLNTTTDAVLYDGLCTDGYEWADWLNMDDPQNDGDWETLRMADMTQACINPKAVEAQRVDSADTTPMVAHISRELGFWCINNEQAGGLCVDYEVRFC
CEKYSAGSDCTEDGYAWTNWLNQDIPFDFGDFETPTYWGEADVCATPIGVEAQPVAGTTGSTEITHIDANEGFWCINEENPSDCADFEARFCCPAVSD
EPDEEGWYTVNEGECTDRIMSWSNWLNGDDPAGEGDYETLAKFNRMDVCGNPTGVQARTRGNDTDAIHLNLESGFWCVNDENTDNGCNDYEVRFCCPR
FRVGSCEGSQASWTGWYNDEWNKKNERLWVSNKQELELNQIYGDGGACKEPTGAEVRVRPTGSTSFQNTMWDYSLNLIQHLDLDGYRCYNDEQLDTAA
GGKYKCVDMEIRFCCEQSLVYGECDQDGYEWSDWLNDDDATGAGDYETLSKFSGKEACAAPIAAQAQAIDDGSVEYTHIDTTIGFYCLNDEQSTGQCA
DFEVRYCCPKMQVGTCNVKGWEWTSYYNIDTAAGTGDWEILKNLEPNQACLNPMGAKVRDTEAGYYGTSDAVTHLSLEGFYCLNEEQPSGMPCADFEI
SYCCPTDETMTCETAEEEWCSENEWCLETRDGPVCKCGDDDFSVDPDSDDFTKYEDGECLANVSPFPAVNGNFTIEYGSCTQYGHVWSSWFNVDTADA
EGDFEILSSLDAHTVCANPTGLRAQALDTGLGAWPVHADLDIGFWCVNGEQDGGQCEDFSIQVCCPVFATCDCPTGHNWTDWYNNDDNLESGDWEMRT
DDMCAKPAALQVRTLDGSPLNNVIHMDNDVGFWCINEENMPEACADYQVSFCCPELEQGECTEYGHTWGSFLDSDDPDGQGDFETLSGFTGYQVCDAP
TGIIAQGINGTDSPDEMKRVSVSEGFVCMNDEFDTCSDFEVQWCCPKWGASANGDDHCMLKGYEWTPWMNEDSPLSGTGDWETIQSQTELKVCSNPTA
IQAFPAGAGATQNTHIDVEIGFWCLNEENSADCADFEVRWCCPTYEDFECDDEGYEWTQWLDRDDPIGEGDYETRFSYPQGSVCEDPTAIQAQARTAG
STSFCHVDLAYGFWCDNSEQPNGAQCADFEVRYCCPKMKEVSCDAEGYGWTVWLDRDDPIESGDWENKDGFPAYIVCKDPLAVEAAVVTGSDGSTAVT
HLDNDQGFWCINDEQPRDEVCADFEVRFCCPNEYTNPCPKPEALVSPNSHIKYNPSNYACECVCDYGYTRDWSTGPPIDASWSCENDEQTCIKMFAPD
TFANAEAHCQAMTGRIVTIPNNDCQNWVDNLDFMNVGYFVLPENVAYSRWGPGHPNDDPGFECQIVDTDQFWRSVDCASHTHAYICQVDQAPITCVKD
VDPCVPTTTECVSFQTCLNADGFLSNDATLGPFAGEPEKVESYQVSCSDPDAIPATNAGGVDGAIHCICDGKSDCYWSSTDFLGALTEEEMCITDTTC
PVSMFKTYTGELQFTRNSFINDPLHNTLQSDLYDFEMAEVLGRIETTALANFESIDWDWSQGHYLVVIWPSTLEANSIQSVYPYGDYVDPGFTSEWGD
GAIQVWETHYNPLVLRDGTVRDIVTSNIEVVLDIRHTKDIADIDELLVFRIAMVPKTWNDKYNDPQTTTVEELSNCLNHMLDDKAKFHPSLRRVGTVS
APKEVAKPVDKGPSGKPNKKKWQKKKKVTKEEWAAKQAKRQARLNGQ
```

Clone G

```
G2a    1 -ASVLSVAAADYACCPYDDYGMPETNCVSVLTEKSPFATADINEALVAQESFACKAWEANAGAANENHKA---ATVNDWGSCGFQRHFP
G2b    1 ------..............................................................---...............
G1     1 L....TI...........Q..V.HSV.SD.........AN.V.--...A..H............L.....DPDNL.H......Y.....

G2a   86 WYNKKPT-AATDPLVKIGLFKATGQDPTYEVFGTGGSYGSTGTNWNMLSLTQGSFSGSAAGNMPITGEVHLGGVCKLFVPVQMDYIRQV
G2b   81 .......-..............................................................................
G1    86 .FHDS.NV.G.ANVNAL...--S.ITAN.Q.L.V...---A..GFS.H...S....ATP.SKQ.V................EF....

G2a  174 SVAGVHINNARIFAGTSQAFSAEMSAVVVGKTAAGTASTPWTGKFHAGTAYCFSVVNIAEFMTNRANLIANANVAGTDTRTAAEEIWYG
G2b  169 .........................................................................................
G1   173 .I...........AG.AVH.N....N...M.DL.D.V.GS.A.....G...................M.............DVNKL.N.

G2a  263 DDQGLQYK----TNGNTESAQFGTIWPGYDGTTDPSPGTAGQKTSGQHKVNMGSNFDVVVHFDSAWCSAHWTHVDMQKDQDHGAGTNTD
G2b  258 ........----.............................................................................
G1   262 .......DQANVNP.AVNP.N.......F....T.TA.VS...NT.........................QL.....VN...N.A---

G2a  348 TAGDNLKGFNNGAGAAKEAMMDAWDKRLPGDVGSCGLCMDHSQVT--PGASIADVTSKCADNTWTAGTDGVTYTMGQTEGFNYHGTPAV
G2b  343 .............................................--.........................................
G1   348 A.TG..L.A.SGTG.-Q............A.T.A....K..ASA.RN.TMHP.T.SAE..T..F...AN.-PH.LSP.DTV..N---.A

G2a  435 PSNCVNI-NEYTSGTDNRWPNAGAWAAFYSFVTCARSDYMIYGFATRTLRTVTLDTPHQENVPASNTYNEAGSATHSDKDNVANSGKQN
G2b  430 .......-.................................................................................
G1   433 VGT.AP.PGL.MPP.TH...........................V....Q.R.EIA.MTES---LTT.PAD.E..KP.TAA.TV--.NR.

G2a  523 NDNRMIVVGGWHQDYRHGQGTCVRFNLRQVGEN----VKYCADGDEEQFNAKDTNSLYKNPTGDTKAWSQDHPNRCTWNWNYKNLANTY
G2b  518 .................................----....................................................
G1   517 ...................TRSHAFVSTFVKS.STSTY.H..S.....E....NAGT..........Q....N..M.........NA.TTNF

G2a  608 DAEAWFDSIDPLEMQVWEVNAVSQTAAEKFIDEPLPGATNTAGFQAAVVANTLTSPVVINVLMQERRRKVNGGTAGTYVDPSNLVMDSS
G2b  603 ..................................G..........................T.........................
G1   606 ........V.........AAG.---.ETVE.V.V.A..TS.S...RN..K.Q....D.T..L.I....Y.R..VD.PA.EG.NS.D..AG

G2a  697 HAVVSALTHFPAICTPG-ADYIAFDNTSGGSTAKITLSCDTAA-KYNGGSAGS---ARMRDHFPDCFFGDELHGQWEWSTG--LSAVDG
G2b  692 ..........DVGDN..-..H....NTS.F.............SP.....T..---.............................--......
G1   692 Y.SS...AS.Q.GNA.STS.WK...T.G..TV.TV..A.K.G--R...VG.SADQH.....E..........I....T.D.ANL.QDL.A

G2a  779 AANTAVWKFWAMLHSH-
G2b  775 ................-
G1   779 .S...........T.TP
```

Clone H

```
MKISAGLLGVALGQNECSAEADTDYTYVYEVDASAAEKSDGYNNGYNSGYNFNQPSYSSQSYDQGYYGRADVGAYELSCWNSNSMRDLNHDNKFAAPG
VNVIGSTTGGMNHQYGYKNSYSAGTKQTAMDYGPNTDLYGSIESGKAIDISDANHSVNAVKPHKWGYQNSNPDAKYHYGHHVNSASSNRGYGPQADTA
YHGFVADDWRYSLRHSGCLYEVKDYTYEATTYDKSSTLVHAAGTGGASVHWVHVFNAHIYPHSDNAINSFRVVMANPVYEGLGYFNFVATYADTAIAA
SGTFTHDPFASTYGTAARYSIAKGTWTLSSTESSWKLEKGATAPAFVLNGGHKSGVAISSFPHNQLGADFRFNVRTLHEFGHGFQDAIHSSDNRADSY
FWYAVDTITITFPHFVSETDSCHAIRSGPVADFSGVDCTEHVHSIIVDENIEVESGAGGANGSPTARLSGAIAVAAIAGGATDGCGGAYGPDGTTNHQ
CASFCQATGVTCGKVLTISNIMSTYDEFHLRQYGTIQEIWAQLQYAYTHTVDKTSTLKTGFESPFPNVFFSAAQIKDIDLSCSCEKCVGYTRSQNMPY
AGDSAVDSRWTNTGITNGDARGDAFWVNNNDN
```

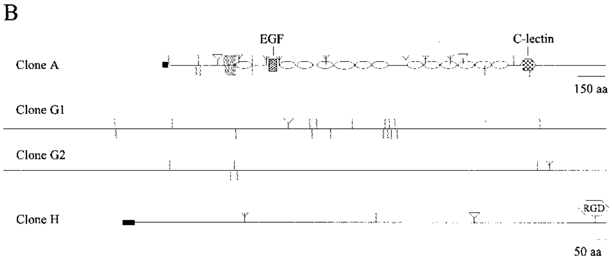

FIG. 4. **Deduced amino acid sequences from cDNAs coding for house polypeptides.** *A*, amino acid sequences are presented for clone A, clones G1, G2a, and G2b, and Clone H. Potential *N*-glycosylation sites are indicated in *bold type* and *underlined*. Potential glycosaminoglycan attachment sites are indicated in *italic type* and *underlined*. Potential RGD cell attachment sequences are indicated in *bold letters* within an *oval*. For sequence comparisons among G clones, *dots* indicate identical residues, and *dashes* indicate gaps. *B*, schematic representations of house proteins. Scaling for Clone A is contracted 3-fold relative to clones G and H. *Vertical wavy lines* indicate potential *O*-glycosylation sites, *tridents* indicate potential *N*-glycosylation sites, *inverted open triangles* indicate potential glycosaminoglycan attachment sites, and *black boxes* indicate predicted secretion signal peptides. For clone A, the cysteine repeat domains are represented by *open ovals*, an epidermal growth factor domain is represented by a *stippled rectangle*, and a C type lectin domain is represented by a *dotted hexagon*. For clone H, the location of the RGD cell attachment sequence is indicated by *RGD* within a *hexagon*. Potential *O*-glycosylation sites were predicted using the NetOGlyc prediction server at the Center for Biological Sequence Analysis. Potential *N*-glycosylation sequences, glycosaminoglycan attachment sequences, and RGD cell attachment sequences were predicted using the PredictProtein server at the University of Columbia. The epidermal growth factor and C-lectin domains were predicted using the PFAM version 5.5 data base at Washington University in St. Louis. Signal peptides were predicted using the SignalP server at the Center for Biological Sequence Analysis, and the PSORT server at the University of Osaka, in combination with data obtained from sequences of

mucins secreted by the epithelia of the respiratory and digestive tracts, and the presence of several conserved cysteine residues in this motif invokes possible roles in the packaging of mucins or in interactions with other components of the mucus (25). Both the mammalian mucins and oikosin 1 are highly glycosylated extracellular proteins, and the degree of conservation among their Cys (sub)domain-like repeats suggests that this domain may have been present in a common ancestor protein prior to the divergence of urochordates and vertebrates.

Expression Analysis—Northern blot analysis (Fig. 6) on poly(A)$^+$ RNA from adult animals using cDNA probes for oikosin 1, oikosin 2A, and oikosin 3 detected single bands for each clone with estimated molecular masses of 7.0, 2.7, and 2.5 kilobases, respectively. These size estimates are in agreement with the full-length cDNAs obtained for oikosins 1 and 3 and indicate that the oikosin 2A cDNA sequence lacks 100–200 bp at the 5′ end of the complete transcript.

The expression patterns of transcripts corresponding to the isolated cDNA clones were assayed by whole mount *in situ* hybridization (Fig. 7). Probes corresponding to different fragments of the nucleotide sequence of oikosin 1 demonstrated that expression of the corresponding gene was restricted to the seven giant cells of the field of Fol (Fig. 7*B*). The Fol region of the oikoplastic epithelium appears to produce different components of the food-concentrating filter, and the seven giant cells are among the largest cells of the epithelium. No signal from any other part of the animal was detected with any probes derived from clone A.

When a nucleotide fragment of oikosin 3 was used as probe, the signal was detected exclusively from a homogeneous group of cells constituting the anterior part of the Fol region (Fig. 7*C*). Probes prepared from oikosin 2A stained several rows of perioral cells in the anterior part of the epithelium, a single row of cells surrounding the anterior Fol region, as well as the three rows of small nasse cells adjacent to the giant cells, and the posterior Fol (Fig. 7, *A* and *A′*). In addition, oikosin 2A was expressed in the chain of pearls and the three central cells of the seven giant cells in the field of Eisen, a region implicated in the synthesis of the inlet filters.

Thus, all of the clones show distinct expression patterns restricted to the oikoplastic epithelium. The patterns are complementary, with no overlap and together cover all cells of the Fol anlage, where the food-concentrating filter is produced. The nuclei of cells in all of these subregions have distinct morphologies that become increasingly elaborate as the animal ages.

DISCUSSION

From small pelagic zooplankton to large baleen whales, filter feeding is an important and recurrent theme in making a living in the ocean. Occupying a phylogenetic position near the transition from invertebrates to vertebrates, the protochordate appendicularians are filter feeders with a circum-global distribution that play a central role in marine ecosystems. The discarded mucous houses, in which they live and filter feed, are a major component of marine snow and represent a significant contribution to global vertical carbon fluxes. However, despite the interesting phylogenetic position and ecological importance of appendicularia, they remain poorly studied.

The apparent association of specific house structures with underlying fields of cells showing diverse and distinctive nuclear morphologies led us to analyze the complexity of proteins composing the house, to clone several cDNAs coding for these proteins, and to examine their spatial expression patterns in

known extracellular proteins obtained from shotgun sequencing of an *O. dioica* cDNA library.

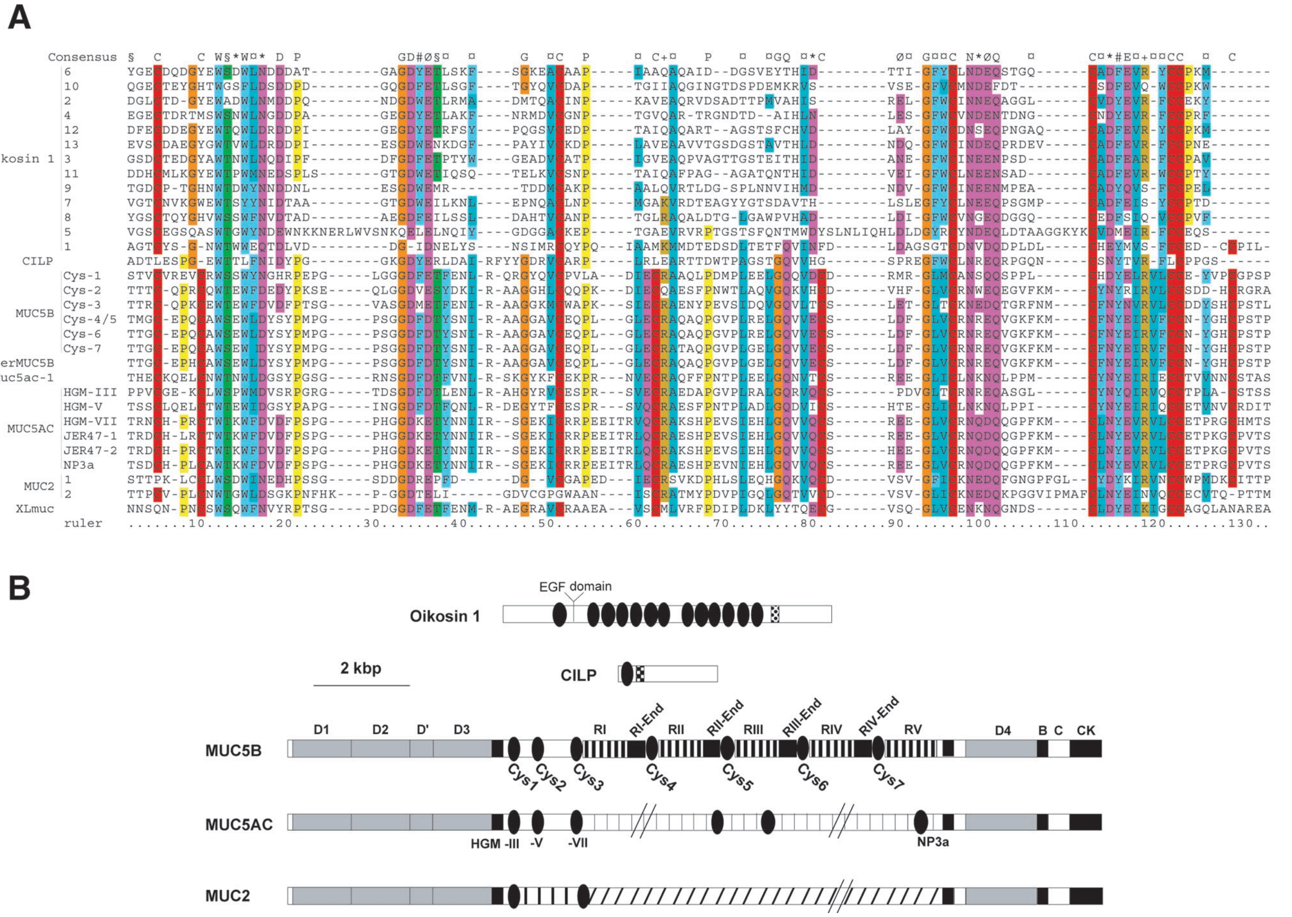

FIG. 5. **Oikosin 1 contains repeats of a motif shared by several vertebrate gel-forming mucins and human CILP.** *A*, clustalX alignment of the 13 repeats of oikosin 1, the Cys subdomains of MUC5B, MUC5AC, MUC2, mouse Muc5ac, cervical MUC5B, a *Xenopus* expressed sequence tag (*Xlmuc*), and a similar motif from CILP. §, hydroxyl residues; #, aromatic residues; ○, hydrophobic residues (including aromatic ones); Ø, acidic residues; *, D, E, N, or Q; +, basic residues. Consensus was assigned to positions where residues with the given property were present in at least 50% of the sequences. Differential coloring indicates highly conserved regions within the repeats. *B*, schematic representation of some of the proteins with Cys subdomain-like motifs (*black ovals*) aligned in *A*. The *dotted rectangle* in oikosin 1 represents the C type lectin domain, and the *checkered rectangle* in CILP represents a type 1 thrombospondin repeat. *Hatched domains* in MUC5B, MUC5AC, and MUC2 represent tandem repeats rich in serine, threonine, and proline of 29, 8, and 23 amino acid residues, respectively. In MUC2 there is an additional imperfect serine-threonine-rich repeat between the two Cys domains. In MUC5B there are five repeats with 11 copies each of the 29-amino acid irregular repeat (*RI*, *RII*, *RIII*, *RIV*, and *RV*). All except RV are followed by a conserved nonrepetitive 111-amino acid sequence (in *black*) called the R-end, also rich in serine, threonine, and proline. The total number of Cys subdomains in MUC5AC is not known. *Diagonal slashes* indicate variations in the inner cores of the mucins. The B, C, CK, and D domains of the mucins are homologous to the corresponding domains of human prepro-von Willebrand factor. Representations of MUC5B, MUC5AC, and MUC2 are adapted from Desseyn *et al.* (25). Schematics of the polypeptides are all drawn to the same scale.

the oikoplastic epithelium. The house is composed of at least 20 polypeptides, a number of which are highly glycosylated, because treatments with combined glycosidases resulted in molecular mass shifts exceeding 100 kDa. This study clearly demonstrated that mRNAs transcribed from genes corresponding to the cloned cDNAs were expressed in very restricted and complementary patterns. Oikosin 1 was restricted to the 14 giant cells in the Fol. The oikosin 2 family was expressed in a more complex pattern, including several rows of peri-oral cells in the anterior epithelium, a single row of cells surrounding the anterior Fol region, and three rows of small nasse cells adjacent to the Fol giant cells and in the posterior Fol. In addition, oikosin 2 was expressed in the chain of pearls and the three central cells of the seven giant cells in the field of Eisen. Oikosin 3 was expressed only from the anterior cells of the Fol region. Both the oikosin 2 family and oikosin 3 are novel proteins showing no homology with any known proteins in data bases. On the other hand oikosin 1 showed limited homology with vertebrate mucins and mammalian CILP.

Oikosin 2 and oikosin 3 are synthesized in regions of the epithelium that are overlaid by mesh structures in the house rudiment (Fig. 1). Oikosin 1 is produced in an intermediate zone between the anterior and posterior mesh zones of the food-concentrating filter. In this respect, the weak sequence homology with human CILP is intriguing. CILP is found only in cartilage and is present in an intermediate layer between collagen mats (26). Homology between oikosin 1 and CILP is restricted to a cysteine-rich domain that had previously only been identified as a subdomain in repeating units found in some vertebrate mucins. The high concentration of CILP in rib cartilage compared with the low levels in tracheal cartilage has been interpreted to suggest that compressive load is a factor in controlling the tissue content of this protein. The fact that CILP is restricted to an interterritorial zone has been taken to indicate that this protein has a structural rather than regulatory role, and it is known that the expression of CILP is increased during the early stages of osteoarthritis (27). One clear difference among the different proteins containing the cysteine-rich domain is that oikosin 1 and the mucins contain multiple repeats, whereas CILP contains only one unit. Tblastn of the oikosin 1 sequence against the entire *Drosophila* and *Caenorhabditis elegans* genomes revealed no homologies.

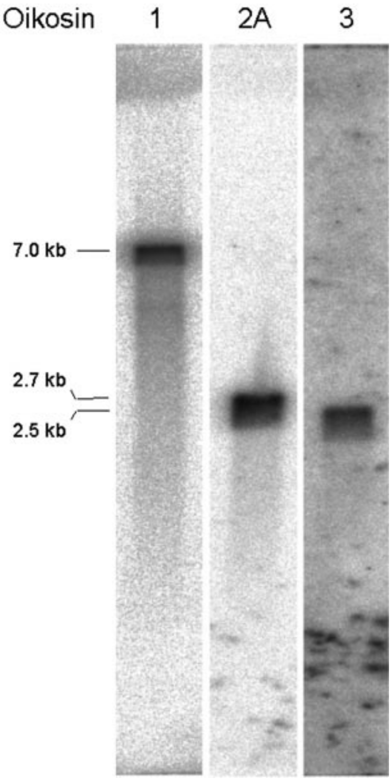

FIG. 6. **Northern blot analysis of poly(A)$^+$ RNA (500 ng/lane) from adult animals probed with different cDNA clones.** *Left lane*, oikosin 1 containing a long reading frame matching all peptides from band A (Fig. 3); *center lane*, oikosin 2A, the longest reading frame matching with peptides from band G (Fig. 3); *right lane*, oikosin 3, the longest reading frame matching with peptides from band H (Fig. 3). The molecular mass of each transcript as estimated by the migration of mammalian rRNA bands is indicated. *kb*, kilobases.

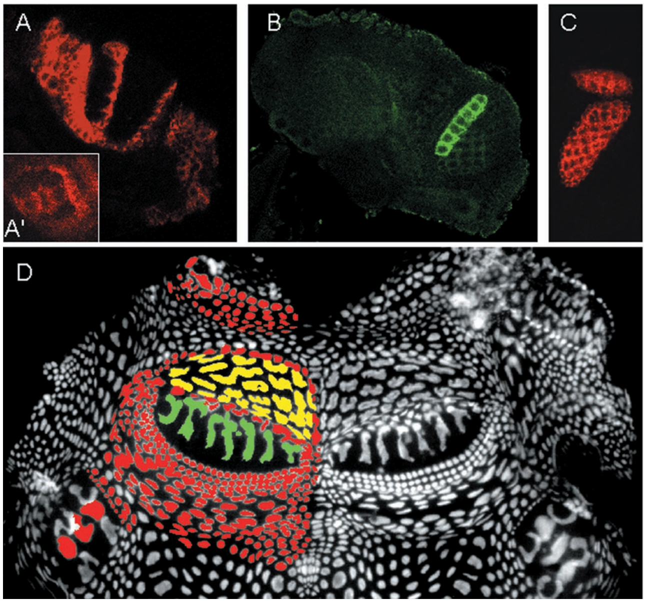

FIG. 7. ***In situ* expression patterns of genes coding for house proteins.** *A–C*, optical sections obtained by confocal microscopy on whole mount fluorescent *in situ* hybridization preparations. The mouth is on the *left*, and the posterior part of the trunk is on the *right*. In *A* and *A′* a fragment of oikosin 2A was used as a probe showing staining on the perioral region and selected regions of the fields of Fol (*A*) and Eisen (*A′*). *A′* is a section through the field of Eisen showing staining of the three central giant cells and of the chain of pearls. In *B* and *C* fragments of oikosin 1 and oikosin 3 stain the seven giant cells and the anterior part of the field of Fol, respectively. *D*, epithelial spread stained with Hoechst 33258, where the expression patterns shown in *A–C* are digitally recreated in different colors. *Red* corresponds to oikosin 2A (*A* and *A′*); *green* corresponds to oikosin 1 (*B*); and yellow corresponds to oikosin 3 (*C*). For simplicity the nuclei of the cells have been digitally colored instead of the cytoplasm. The mouth is at the *top*, and the posterior part of the epithelium is at the *bottom*.

Therefore, thus far, we are only able to trace the origins of this domain to a common ancestor protein present prior to the divergence of urochordates and vertebrates. The conservation of this domain in organisms as distant as humans and appendicularians suggests that it is sufficiently important to merit further study.

We have shown that the easily accessible oikoplastic epithelium of *O. dioica* serves as a unique template for the synthesis of different house components from defined cellular regions with distinct nuclear morphologies. Therefore, we now possess markers to study the coordinate regulation of gene expression in the house building process, cell-cell interactions implicated in pattern formation, and the role that nuclear architecture might play in regulating gene expression. Superficially, this system resembles synthesis of the insect cuticle from the underlying epidermis or the production of the egg chorion from follicle cells in *Drosophila*. These processes have already been the targets of some experimental investigation, and the amplification of chorion genes has been partially characterized (28). However, neither of the latter two systems presents the same compelling association of specific structures with underlying subregions of cells. The insect cuticle contains a chitinous assembly zone directly above the epidermal microvilli and the perimicrovillar space. During intermoult, the epidermis secretes peptides constitutively, and these traverse the perimicrovillar space and form lamellae in the assembly zone (29). Furthermore, some components of the cuticle are produced in other organs and transported to the cuticle via the hemolymph (30). In *Drosophila*, the composition of the egg chorion appears relatively uniform, although there are some dorsal anterior cells over the oocyte that synthesize specialized chorion structures (28). Therefore, the oikoplastic epithelium of *O. dioica* offers significant advantages when compared with alternative model systems and seems exceptionally well suited to probing the fundamental questions raised above.

Acknowledgment—We thank Rein Aasland (University of Bergen) for assistance with the Cys domain alignments.

REFERENCES

1. Wada, H., and Satoh, N. (1994) *Proc. Natl. Acad. Sci. U. S. A.* **91,** 1801–1804
2. Christen, R., and Braconnot, J.-C. (1998) in *The Biology of Pelagic Tunicates* (Bone, Q., ed), pp. 265–271, Oxford University Press, New York
3. Alldredge, A. (1977) *J. Zool.* **181,** 175–188
4. Flood, P. (1991) *Mar. Biol.* (*N.Y.*) **111,** 95–111
5. Flood, P., and Deibel, D. (1998) in *The Biology of Pelagic Tunicates* (Bone, Q., ed), pp. 105–124, Oxford University Press, New York
6. Lohmann, H. (1899) *Schr. Naturwiss. Ver. Schleswig-Holstein* **11,** 347–407
7. Martini, E. (1909) *Z. Wiss. Zool.* **94,** 563–626
8. Lohmann, H., and Bückmann, A. (1926) in *Deutsche Südpolar-Expedition 1901–1903*, Vol. 18 (Zoologie 10), pp. 63–221, Polarphilazelie e.V., Leverkusen, Germany
9. Fenaux, R. (1985) *Bull. Mar. Sci.* **37,** 498–503
10. Fenaux, R. (1971) *Z. Morphol. Tiere* **69,** 184–200
11. Lohmann, H. (1933) in *Handbuch der Zoologie* (Kükenthal, W., and Krumbach, T., eds) Vol. 5, pp. 15–164, W. de Gruyter, Berlin
12. Shägger, H., and von Jagow, G. (1987) *Anal. Biochem.* **166,** 368–379
13. Møller, H., and Poulsen, J. (1995) *Anal. Biochem.* **226,** 371–374
14. Shevchenko, A., Wilm, M., Vorm, O., and Mann, M. (1996) *Anal. Chem.* **68,** 850–858
15. Wilm, M., and Mann, M. (1996) *Anal. Chem.* **68,** 1–8
16. Shevchenko, A., Chernuschevich, I., Ens, W., Standing, K., Thomson, B., Wilm, M., and Mann, M. (1997) *Rapid Commun. Mass Spectrom.* **11,** 1015–1024
17. Jay, G., Culp, D., and Jahnke, M. (1990) *Anal. Biochem.* **185,** 324–330
18. Nakai, K., and Kanehisa, M. (1992) *Genomics* **14,** 897–911
19. Nielsen, H., Engelbrecht, J., Brunak, S., and von Heijne, G. (1997) *Protein Eng.* **10,** 1–6
20. Desseyn, J.-L., Guyonnet-Duperat, V., Porchet, N., Aubert, J.-P., and Laine, A. (1997) *J. Biol. Chem.* **272,** 3168–3178
21. Buisine, M.-P., Desseyn, J.-L., Porchet, N., Degand, P., Laine, A., and Aubert, J.-P. (1998) *Biochem. J.* **332,** 729–738

22. Desseyn, J.-L., Buisine, M.-P., Porchet, N., Aubert, J.-P., Degand, P., and Laine, A. (1998) *J. Mol. Evol.* **46,** 102–106
23. Pilar, L., Neame, P., Sommarin, Y., and Heinegård, D. (1998) *J. Biol. Chem.* **273,** 23469–23475
24. Krieg, J., Hartmann, S., Vicentini, A., Glasner, W., Hess, D., and Hofsteegne, J. (1998) *Mol. Biol. Cell* **9,** 301–309
25. Desseyn, J.-L., Aubert, J.-P., Porchet, N., and Laine, A. (2000) *Mol. Biol. Evol.* **17,** 1175–1184
26. Lorenzo, P., Bayliss, M., and Heinegård, D. (1998) *J. Biol. Chem.* **273,** 23463–23468
27. Bayliss, M., Venn, M., Maroudas, A., and Ali, S. (1983) *Biochem. J.* **209,** 387–400
28. Calvi, B., Lilly, M., and Spradling, A. (1998) *Genes Dev.* **12,** 734–744
29. Locke, M., Kiss, A., and Sass, M. (1994) *Tissue Cell* **26,** 707–734
30. Csikós, G., Molnár, K., Borhegyi, N., Talián, G., and Sass, M. (1999) *J. Cell Sci.* **112,** 2113–2124