ABSTRACT We have deduced the 1675-amino acid sequence of rat clathrin heavy chain from cDNA clones and predict a protein of Mr 191,569. We have established the polarity of the heavy chain and assigned sequence positions to several structural landmarks of the clathrin leg. The terminal domain at the distal end of the clathrin leg is at the amino terminus of the heavy chain. It is connected to the distal segment by a flexible "link" from Tyr-479 to Arg-523. There is an unusual sequence at the carboxyl terminus that may form the globular projection at the vertex of the clathrin trimer. We suggest that a possible site of heavy-chain–light-chain interaction is located in the proximal segment. Comparison with other partially sequenced mammalian clathrin heavy chains shows that the primary structure is highly conserved. The heavy chain is unrelated to other classes of structural proteins.

Coated pits and coated vesicles are organelles that mediate vesicular intracellular transfer between membranous compartments (reviewed in refs. 1–3). Clathrin is the major protein constituent of the coat that surrounds the cytoplasmic face of these structures (3–5). It is believed that the assembly of the coat provides the driving force that leads to the specific entrapment of receptors and other macromolecules on the coated pits and to their vesiculation into coated vesicles (3).

Clathrin is a three-legged structure of pinwheel shape (6, 7) composed of three heavy chains (M(), 191,569) and three light chains (M(), 23–25,000, deduced from the sequence of rat (8) and bovine (9) cDNA). Each leg contains a single heavy chain and a noncovalently bound light chain (6, 7). The three heavy chains are held together at the center of the trimer (skelton) through noncovalent interactions (6, 7); the presence of light chains is not required for its trimeric integrity (7). Each leg is ~500 Å long and has a defined substructure (6, 7, 10–12). Cleavage experiments combined with electron microscopy show that the heavy chain runs from one end of a leg to the other end (10). The structural divisions of the leg must, therefore, correspond to contiguous parts of the heavy-chain sequence. The principal subdivisions of the leg are the terminal domain and the distal and proximal segments. The terminal domain is a globular region of M(), ~50,000, located at the distal end and connected to the distal segment by a proteolytically sensitive "link" region. The distal segment is a linear section, ~255 Å long, with several sites that are particularly sensitive to enzymatic cleavage. The proximal segment, ~170 Å long, contains the light-chain binding site (13–15). The sharp bend between the proximal and distal segments is referred to as the "joint." A small domain at the vertex (12) probably mediates trimer interactions. The proximal and distal segments are probably responsible for most of the interactions that hold together a cage (10, 13, 16).

We report here the complete amino acid sequence of the heavy chain of clathrin deduced from rat brain cDNA clones. This information has allowed us to determine the orientation of the heavy chain with respect to the leg and to assign several structural landmarks of the leg with the primary structure of the heavy chain. We suggest a possible light-chain binding site, and report the partial amino acid sequence of bovine clathrin heavy chain and the extensive sequence similarities with the rat homolog.

METHODS

Clathrin. Clathrin was purified from calf brain coated vesicles by either the urea or the Tris-dissociation method (10, 17). Heavy and light chains were separated using KSCN (18) or by NaDodSO4/polyacrylamide gel electrophoresis (19).

Proteolytic Cleavages, Purification of Peptides, and Amino-Terminal Sequencing. The CNBr-digestion protocol was as follows: ~20 mg (at 4 mg/ml) of purified clathrin was dialyzed against ammonium bicarbonate, pH 8.0/0.5 mM dithiothreitol/0.02% NaN() and mixed with 16 ml of 88% (vol/vol) formic acid and 4 ml of a freshly made solution of 100 mg of CNBr per ml of 88% (vol/vol) formic acid for 24 hr in the dark at room temperature. After lyophilization, the sample was resuspended in 0.6 ml containing 2% (wt/vol) NaDodSO4 and 5 mM dithiothreitol, for 10 min at 80°C. At this point, 1.2 ml of 80 mM Tris-HCl, pH 8.9/0.2 M urea (freshly made) was added, and, after 5 min at room temperature, the sample was loaded onto a preparative NaDodSO4/15% polyacrylamide gel (19). Proteolytic fragments of known location along the clathrin leg were obtained by enzymatic proteolysis of clathrin cages (10) followed by NaDodSO4/12% polyacrylamide gel electrophoresis. The bands were identified with brief Coomassie brilliant blue or 1 M KCl staining, and selected peptides were isolated by electroelution into 0.1% NaDodSO4/25 mM ammonium bicarbonate followed by extensive dialysis against 0.1% NaDodSO4. Sample volume was reduced to ~50 µl in a SpeedVac, and, after centrifugation for 10 min in an Eppendorf centrifuge, the supernatant was loaded directly into a gas-phase Applied Biosystems amino acid sequencer.

cDNA Isolation. A A10 rat brain cytoplasmic poly(A)+ cDNA library (20) was screened with a unique-sequence 122-base-pair DNA probe (Fig. 1 Right) as described (22). Additional clones were isolated from A10 cDNA libraries constructed from rat liver cytoplasmic poly(A)+ RNA, from bovine brain poly(A)+ RNA (20), and from a plasmid cDNA.

1To whom reprint requests should be addressed.

2This sequence is being deposited in the EMBL/GenBank data base (Bolt, Beranek, and Newman Laboratories, Cambridge, MA, and Eur. Mol. Biol. Lab., Heidelberg) (accession no. J03583).
were amide gel chains) was of oligonucleotide contaminating acid purified (nucleotides probe and was (21) ratio of yield of amino acid sequence clathrin heavy chains (data not shown) was performed; amino acid sequence determination of the M_r = 23,000 fragment yielded two sequences corresponding to a mixture of peptide 1 and a contaminating peptide 3 (also from heavy chain, Fig. 2) with a molar ratio of ≈ 2:1. Size markers are indicated. (Right) The unique sequence 122-base-pair DNA probe used in the first cDNA library search was derived from the amino acid sequence of peptide 1. The probe was designed according to the pattern of codon usage in rat (21) and was assembled by annealing two partially complementary 67-mers (Upper and Lower) followed by a fill-in reaction using 32P-labeled nucleotides with Klenow fragment of DNA polymerase I. A nucleotide identity of 82% with the clathrin heavy-chain cDNA was obtained (nucleotides 1051–1173, Fig. 2).

library from rat brain cerebral cortex poly(A)^+ RNA (gift of M. Brownstein, National Institute of Mental Health). These searches were carried out at high stringency using restriction endonuclease fragments from rat clathrin heavy-chain cDNA clones. The hybridizing fragments were subcloned into plasmid vectors.

DNA Sequencing and DNA and RNA Blots. DNA sequences were obtained on both strands (23). Nucleotide positions 1427–1447 were sequenced on only one strand; ≈ 70% of the sequence including the 5' and 3'-untranslated regions was derived from two or more independent clones. RNA and DNA blots were done as described elsewhere (24, 25).

RESULTS AND DISCUSSION

Cloning and the Sequence of Clathrin Heavy Chain.

The amino terminus of bovine clathrin heavy chain appears to be blocked regardless of how the clathrin is purified. Thus, to obtain internal amino acid sequence information, peptides from bovine clathrin were generated by partial cleavage with CNBr. A M_r = 23,000 fragment yielded a sequence of 51 amino acids (peptide 1, Figs. 1 Left and 2) from which we selected a region of 41 consecutive amino acids to design a probe spanning 122 nucleotides of unique sequence (Fig. 1 Right). Six independent and related hybridizing clones were identified after initial screening of a λgt10 library from rat brain. The deduced amino acid sequence of the smallest clone confirmed its identity as a partial clathrin heavy-chain cDNA. The agreement included the complete amino acid sequence of peptide 1 as well as 26 out of 27 amino acids of peptide 2 (Figs. 1 Left and 2). Additional recombinants were isolated from a search of the rat brain λgt10 cDNA library, from the rat liver λgt10 cDNA library, and from the rat brain plasmid cDNA library. The compiled DNA sequence for clathrin heavy-chain cDNA and the deduced complete amino acid sequence are given in Fig. 2. The nucleotide sequence spans 6074 base pairs, in good agreement with the mRNA size of ≈ 6000 nucleotides predicted by RNA blot analysis (Fig. 3A). The first ATG start codon in the open reading frame, defined as position 1, is included in the sequence GCCATGG, which fits the criteria for an eukaryotic consensus translation initiation site (26). The translational stop codon at position 5026 defines an open reading frame that predicts a clathrin heavy chain of 1675 amino acids (M_r = 191,569). The stop codon is followed by a 3'-untranslated region of 949 base pairs, including two putative polyadenylation signals (AATAAA and AT- TAAA) at positions 5097 and 5953 and a poly(A) tail.

The molecular weight of clathrin heavy chain determined from the cDNA is slightly higher than the values reported on the basis of NaDodSO4/polyacrylamide gel electrophoresis (M_r = 180,000) (4) or equilibrium sedimentation in 6.5 M guanidinium hydrochloride (M_r = 170,000) (27). We believe that we can rule out the use of a more distal start codon or a posttranslational cleavage of the amino terminus, because our assignment of the start codon is consistent with the size of several proteolytic fragments generated by enzymatic cleavages near the amino terminus of the heavy chain (see below). We also can rule out extensive posttranslational processing of the carboxyl terminus, since the start of one of our CNBr-generated fragments (peptide 3, Fig. 2) lies only 72 amino acids away from the carboxyl end of the molecule.

A computer search (28) with the heavy-chain amino acid sequence showed no significant homology with clathrin light chains (8) nor with other sequenced proteins that have been published in the National Biomedical Research Foundation protein data bank. This result suggests that clathrin heavy chain is a member of a distinct class of structural proteins.

RNA blot analysis indicates that a single mRNA species is present in brain, heart, kidney, liver, lung, and testis (Fig. 3A). DNA blot analysis of rat genomic DNA performed at high stringency using a cDNA probe from the coding region (data not shown) or from the 3'-untranslated region indicates the absence of pseudogenes and suggests that only one gene codes for clathrin (Fig. 3B). At this point we cannot rule out the existence of related genes with low DNA homology.

Primary Structure and Orientation of the Heavy Chain in the Leg of a Clathrin Trimer. The determination of the complete amino acid sequence of the rat heavy chain, as well as knowledge of the relationship between morphological domains and proteolytic fragments and of the amino-terminal sequences of selected fragments, has allowed us to establish the polarity of the heavy chain and to assign sequence position to some regions of the molecule (Fig. 4). The terminal domain is a compactly folded structure, comprising 1/3 of the total heavy-chain mass, located at the distal end of a leg (10). Terminal domains, generated by enzymatic proteolysis of bovine clathrin cages and purified by preparative NaDodSO4/12% polyacrylamide gel electrophoresis, are completely resistant to amino-terminal sequence determination. A blocked amino terminus is found, whatever the protease used to release terminal domain (chymotrypsin, subtilisin, or thermolysin), and the block is thus independent of the precise location of the cleavage point between the terminal domain and the rest of the clathrin molecule (10). Since the amino terminus of the intact heavy chain is also blocked, we conclude that the terminal domain, the distal end of the clathrin leg, lies at the amino terminus of the heavy chain (Fig. 4).

When terminal domain is produced by chymotryptic proteolysis followed by gel filtration under nondenaturing conditions (10), we can detect an unblocked amino terminus

Fig. 2. Complete nucleotide and deduced amino acid sequence of the heavy chain of clathrin from rat brain. Amino acids (single-letter code) are numbered from the assigned initiation codon. The open reading frame contains 1675 amino acids. Two putative polyadenylation signals (AAATAAA and ATTAAA, nucleotides 5907 and 5953) and an upstream in-frame translational stop codon (nucleotide ~ 33) are underlined. The segments of predicted rat amino acid sequence that correspond to the sequence of the bovine brain heavy-chain proteolytic fragments are underlined (peptides 1-8); X denotes amino acids whose identity by Edman degradation was ambiguous. The comparison with the predicted amino acid sequence of bovine clathrin heavy chain (segments 9 and 10 at amino acid positions 403-509 and 767-937) deduced from a cDNA clone is underlined. The possible binding site to the light chains (heptads 1-11) is indicated by broken lines.

following Tyr-479 (peptide 4, Fig. 2). This cleavage site defines a fragment of $M_0$ 53,300 (measured from the amino terminus of the predicted rat heavy-chain sequence). Its size agrees closely with that of terminal domain generated by chymotryptic digestion of bovine heavy chain (10), as determined by analytical NaDodSO4/polyacrylamide gel electrophoresis. We believe that the sequence of peptide 4 derives from a small fragment, generated by chymotryptic cleavage, that remains attached during purification to the principal part of the terminal domain. We note that this observation is also
consistent with the assignment of the terminal domain to the amino terminus of the heavy chain.

Limited digestion of cages with chymotrypsin or trypsin produces fragments of $M_r = 105,000$ corresponding to cleavages in the distal segment at $\sim 315$ Å from the center of the clathrin trimer (10, 13, 16) (Fig. 4). Comparison of their amino-terminal sequences with the predicted sequence from rat cDNA clones shows that the cleavage sites for chymotrypsin and trypsin map to the carboxyl side of Tyr-634 and Arg-638 (peptides 5 and 6, Fig. 2), respectively. The proximity of the target sites suggest that they may be located on a relatively exposed section, perhaps at the boundary between two more compact domains. The positions just described predict fragments of $M_r = 120,000$ measured from the carboxyl terminus of the molecule, corroborating our assignment for the relative orientation of the clathrin leg and the heavy-chain primary structure (Fig. 4).

Another product obtained from limited digestion of bovine clathrin cages is a fragment of $M_r = 116,000$ (as measured by gel electrophoresis), which corresponds to a cleavage in the distal segment, closer to the terminal domain than the cleavage producing $M_r = 105,000$ (10). The $M_r = 116,000$ fragment obtained after trypsin digestion contains a mixture of two peptides whose amino-terminal sequences map to the carboxyl side of Lys-507 and Arg-523 (peptides 7 and 8, Fig. 2). These sites thus predict complementary fragments of $M_r = 58,000-59,000$, measured from the amino terminus of the heavy chain. Certain terminal-domain species are indeed of just this size. For example, when thermolysin is used to digest clathrin cages, a fragment of $M_r = 59,000$ (10) and Stokes’ radius of $\sim 50$ Å (T.K., unpublished data) appears after brief treatment. Longer digestion produces the terminal domain proper [$M_r = 52,000$; Stokes’ radius $\sim 35$ Å (10)]. Both fragments have blocked amino termini and similar peptide maps (T.K., unpublished data). Thus, the cleavage sites that define the amino end of the $M_r = 116,000$ fragment and the carboxyl end of the $M_r = 59,000$ fragments must effectively coincide, at a position $\sim 400$ Å from the center of the trimer (see length measurements in ref. 10). We also conclude that the sequence following Tyr-479 (the boundary of the terminal domain proper) and reaching Arg-523 corresponds to the link between terminal domain and distal segment. This linking sequence probably spans $\sim 30$ Å to account for the difference in Stokes’ radii of the $M_r = 59,000$ and $M_r = 52,000$ fragments. We note that, in the assembled cage, the link lies under a vertex and allows the terminal domain to project inward toward the membrane (10, 29, 30).

Another structural landmark is a projection seen by rapid-freeze, deep-etch electron microscopy at the center of the trimer on its upper or cytoplasmic side (12). This may be a specialized structure involved in the noncovalent interactions holding the three heavy chains together. Inspection of the heavy-chain primary sequence shows that all amino acids are distributed evenly over the length of the heavy chain with the exception of the carboxyl-terminal 47 amino acids, where 10 prolines, 8 glycines, no charged amino acids, and a repeat of the sequence Tyr-Gly-Gln-Pro-Gln at amino acid positions 1634 and 1664 are found. We propose that this region is involved in trimer formation. Although the assignment of this unusual sequence to the trimer vertex is hypothetical, it agrees with the overall linear disposition of the heavy chain along the clathrin leg. It is also consistent with the observation that cleavage very near the carboxyl terminus removes the corresponding leg from the trimer (see figure 4h of ref. 10). We note that the noncovalent trimeric interactions that hold together virus assemblies are mediated by intertwined polypeptide chains rather than by surface contacts of the subunits (31, 32).

Finally, we have searched in the amino acid sequence for elements that could correspond to a concatenation of similar simple domains that could account for the elongated shape of the leg. The analysis, performed by Fourier transformation (33) and by dot-matrix homology comparison (34), has not revealed any simple repeat pattern. A similar lack of repeating elements has frustrated understanding of the highly elongated tail fiber of T4 phage (35). We do find, mostly along the proximal segment, the sequences from residue 360 Ala-Glu-Glu-Leu-Phe, from residue 450 Ser-Glu-Glu-Leu-Gly, from residue 1056 Phe-Glu-Glu-Ala-Phe, from residue 1150 Trp-Glu-Glu-Leu-Val, from residue 1186 Leu-Glu-Glu-Phe-Ile, from residue 1283 Leu-Glu-Glu-Leu-Ile, from residue 1296 Phe-Glu-Glu-Leu-Ile, and from residue 1550 Ala-

![Fig. 3. RNA and DNA blot analysis. (A) RNA blot analysis. Poly(A)"RNA (1 μg) from rat brain (lane B), heart (lane H), kidney (lane K), liver (lane L), lung (lane LU), and testis (lane T) were probed with heavy-chain cDNA spanning nucleotides 4237-4426 of coding region. All tissues contain a mRNA species of similar size (≈6000 nucleotides). β-Actin mRNA was used as internal reference. 28S and 18S RNA size markers are indicated. (B) Southern blot analysis. Rat liver genomic DNA was digested with EcoRI (lane 1), Pvu II (lane 2), BamHI (lane 3), or Kpn I (lane 4) and probed with heavy-chain cDNA spanning nucleotides 3332-5842 of 3'-untranslated region (B). Size markers (arrowheads) were 23, 130, 9416, 6557, 4361, 2322, and 2027.

![Fig. 4. Relationships between the primary structure, structural domains, and amino-terminal sequence of selected fragments of clathrin heavy chain. Each leg of a clathrin trimer (triskelion) contains one heavy chain and one tightly bound light chain. The terminal domain lies at the amino terminus of the heavy chain whereas the vertex probably contains the carboxyl end of the molecule. Assignment of amino-terminal sequence of selected proteolytic fragments is indicated. Boxed sequences refer to enzymatic cleavage sites accessible on the native molecule. Sequences are in the single-letter amino acid code.](image-url)
Glu-Glu-Leu-Leu. The structural or functional significance of this motif (if any) is obscure.

Possible Site of Heavy-Chain–Light-Chain Interaction. Clathrin light chains are extended molecules that bind tightly to the proximal segment of each heavy chain (13–15). Analysis of the primary structure of light chains A and B has indicated the presence on each molecule of a pattern of 10 heptad motifs (8). Each motif is of the form a, b, c, d, e, f, and g, where positions a and d denote positions that are in general occupied by neutral or hydrophobic amino acids. This pattern is a signature for a coiled-coil α-helical conformation (33, 36, 37). Since this configuration requires close interaction with another polypeptide chain, we postulated that the heptad mediates binding to heavy chain (8).

We have found one region of the heavy chain that displays a weak pattern of heptad repeats. It starts at Val-1107, spans 11 consecutive heptads with positions a and d occupied by uncharged amino acids, and maps to the proximal segment of the clathrin leg. There is a “skip residue” (Gln-1142) between the fifth and sixth heptads, as in the binding region of all light chains. Other aspects of this candidate sequence make a coiled-coil picture less compelling, however. At the end of the fourth heptad lies Pro-1134, an amino acid that weakens or disrupts α-helices. Moreover, one-dimensional Fourier analysis of the segment does not provide evidence for a strong repeat with period 7/2 despite the prevalence of uncharged residues in positions a and d.

The Primary Structure Is Highly Conserved. We have found that the primary structure of rat and bovine clathrin heavy chains are remarkably similar. As indicated in Fig. 2, all the bovine peptide sequences that were determined as well as the predicted amino acid sequence derived from a bovine heavy-chain cDNA clone are virtually identical to the predicted rat sequence (396 out of 401 amino acids compared). We have obtained similar results in a comparison with the predicted amino acid sequence deduced from a human clathrin heavy-chain cDNA clone (99% identity over 122 amino acids; S. Frucht and T.K., unpublished observations). In this regard a partial sequence for the yeast clathrin heavy chain, derived from genomic DNA clones (38, 39) indicates >50% identity with the higher eukaryote counterparts (>30% of the chain has been sequenced so far); S. Lemmon, personal communication; T. Hasson, G. Payne, and R. Schekman, personal communication; T.K. and J.B., clone provided by G. Payne and R. Schekman). Perhaps not surprisingly, the amino acid sequences for clathrin light chains obtained from rat (8), bovine (ref. 9; T.K. and J.B., unpublished data), and human cDNA (S. Frucht and T.K., unpublished data) also display a high level of sequence similarity among themselves. The essentially identical primary structure of mammalian clathrin heavy chains may be the result of a selective pressure to maintain the large number of surface interactions that must be accommodated when clathrin assembles into the lattice of coated pits and coated vesicles.

We thank A. Kirsch for help in protein purification; D. Andrews and B. Lane of the microprocessing facility, Harvard Biobots, for peptide sequencing; J. J. Monroe, A. Ahn, M. Laverty, V. Gregg, and L. Arthur for help at various stages of cloning and sequencing; R. Tizard for advice on DNA sequencing; M. Brownstein (National Institutes of Mental Health) for the kind gift of the rat brain plasmid library; U. Nudel for the β-actin probe; W. Gilbert for computer time; L. Pops for computer system support; I. Mocchetti and K. Mallin for useful discussions; C. Beepot and L. M. Oliphant for secretarial assistance. This project was supported in part by Grant R01 GM 36548-01 from the National Institutes of Health (T.K.); Grant DMB WS-02920 from the National Science Foundation (S.C.H.); Grant MH 38819 from the National Institute of Mental Health (J.B.); and by a contribution from the Foundation for Medical Research (T.K.). T.K. is an Established Investigator of the American Heart Association, and J.B. is recipient of an Irma H. Hirsch Career Scientist Award.