Proc. Natl. Acad. Sci. USA
Vol. 87, pp. 955–959, February 1990
Biochemistry

Molecular cloning of human protein 4.2: A major component of the erythrocyte membrane

(band 4.2/membrane skeleton/factor XIII/transglutaminase/cDNA)

LANPING AMY SUNG*†, SHU CHIEN*, LONG-SHENG CHANG‡, KAREL LAMBERT*, SUSAN A. BLISS‡, ERIC E. BOUHASSIRA§, RONALD L. NAGEL§, ROBERT S. SCHWARTZ§, AND ANNE C. RYBICKI§

*Departments of Applied Mechanics and Engineering Sciences-Bioengineering and Medicine, and Center for Molecular Genetics, University of California at San Diego, La Jolla, CA 92093; ‡Department of Biology, Princeton University, Princeton, NJ 08544; and §Division of Hematology, Albert Einstein College of Medicine/Montefiore Medical Center, Bronx, NY 10467

Communicated by Russell F. Doolittle, November 16, 1989

ABSTRACT **Protein 4.2 (P4.2) comprises ≈5% of the protein mass of human erythrocyte (RBC) membranes. Anemia occurs in patients with RBCs deficient in P4.2, suggesting a role for this protein in maintaining RBC stability and integrity. We now report the molecular cloning and characterization of human RBC P4.2 cDNAs. By immunoscreening a human reticulocyte cDNA library and by using the polymerase chain reaction, two cDNA sequences of 2.4 and 2.5 kilobases (kb) were obtained. These cDNAs differ only by a 90-base-pair insert in the longer isoform located three codons downstream from the putative initiation site. The 2.4- and 2.5-kb cDNAs predict proteins of ≈77 and ≈80 kDa, respectively, and the authenticity was confirmed by sequence identity with 46 amino acids of three cyanogen bromide-cleaved peptides of P4.2. Northern blot analysis detected a major 2.4-kb RNA species in reticulocytes. Isolation of two P4.2 cDNAs implies existence of specific regulation of P4.2 expression in human RBCs. Human RBC P4.2 has significant homology with human factor XIII subunit a and guinea pig liver transglutaminase. Sequence alignment of P4.2 with these two transglutaminases, however, revealed that P4.2 lacks the critical cysteine residue required for the enzymatic crosslinking of substrates.**

Membrane skeletal proteins play an important role in regulating the viability and mechanical properties of erythrocytes (RBCs) (1, 2). All major RBC membrane proteins have been identified, most have been characterized, and several have been cloned (3–6). Isoforms of many RBC membrane proteins have been identified in various types of cells and tissues (1, 2). Protein 4.2 (P4.2), which represents ≈5% of the protein mass of human RBC membranes, is one of the last major membrane proteins to be characterized. P4.2 has an apparent molecular mass of ≈72 kDa and associates with the cytoplasmic domain of the anion exchanger, band 3 (7). Recent evidence suggests that P4.2 interacts with ankyrin and may function to stabilize ankyrin in the membrane (8). Individuals whose RBCs are severely deficient in P4.2 experience various levels of anemia, further indicating an important functional role for this protein (8).

We now report the molecular cloning and characterization of the full-length cDNA for human RBC P4.2.¶ Two cDNA sequences have been identified that differ only by a 90-base-pair (bp) insert located near the 5′ end of the coding region. The presence of two P4.2 cDNAs resembles the transcript heterogeneity found in membrane skeletal protein 4.1 (9–11) and nonerythroid α-spectrin (6) and suggests that regulation of P4.2 expression exists in human RBCs, possibly by alternative splicing.

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

MATERIALS AND METHODS

Screening of λgt11 cDNA Library. Affinity-purified rabbit anti-human P4.2 IgG prepared by Rybicki *et al.* (8) was used to screen a cDNA expression library in λgt11 constructed from human reticulocyte mRNA, kindly provided by J. G. Conboy and Y. W. Kan (5). Immunoscreening of the λgt11 expression library was performed according to Huynh *et al.* (12), except that positive clones were identified with goat anti-rabbit IgG conjugated with horseradish peroxidase (Bio-Rad). For each 150-mm Petri dish, 5×10^4 plaque-forming units were used.

Subcloning and Sequence Analysis. cDNA inserts from positive phage clones were subcloned into pBS(+) plasmids (Stratagene). Unidirectional deletion clones were generated by using BAL-31 exonuclease (13), and cDNA fragments were sequenced with T3 and T7 primers by the dideoxynucleotide chain-termination method (14). Sequence analysis and GenBank data base searches were performed by IBI Pustell sequence analysis software (International Biotechnologies).

5′-End Extension of cDNA. Three oligonucleotides were prepared in a technique based on the polymerase chain reaction (PCR) to synthesize the missing 5′ sequence of the partial cDNA clone: p1 was composed of nucleotides (nt) 7–23 of clone 7 (c.7); p2 was complementary to nt 36–52 of c.7 plus an *Eco*RI restriction site at its 5′ end; p3 was composed of the *Eco*RI polylinker and the poly(dC) originally used for the first-strand cDNA synthesis when the library was constructed. The sequences of p1, p2, and p3 were, respectively, 5′-dTGAGGATGCTGTGTTCC-3′, 5′-dTCGAATTCGTACTCCATGCGCTGAG-3′, and 5′-dGCGGAATTCCCCCCCCCCCCCC-3′, with the *Eco*RI sites underlined. p2 and p3 were used as PCR primers, and the reticulocyte cDNA library (5 μl with 10^6 phages per μl) was used as a template. The reaction product was electrophoresed and stained with ethidium bromide. The major band was excised and subcloned into pGEM 3zf plasmids (Promega). From >500 transformants, 24 colonies were randomly chosen to make minipreparations of plasmid DNA, of which 80% were positive when hybridized with ^{32}P-labeled p1.

Western Blot Analysis of Fusion Proteins. Recombinant lysogens of three positive λgt11 clones were prepared and induced for expression of β-galactosidase fusion proteins (12). Clear lysate containing the fusion proteins was separated by SDS/PAGE (7.5% polyacrylamide gel) (15), transferred to

Abbreviations: LTG, guinea pig liver transglutaminase; nt, nucleotide(s); ORF, open reading frame; PCR, polymerase chain reaction; P4.2, protein 4.2 of human erythrocytes; RBC, erythrocyte; $XIII_a$, subunit a of human factor XIII.
†To whom reprint requests should be addressed.
¶The sequences reported in this paper have been deposited in the GenBank data base (accession nos. M30646 and M30647).

nitrocellulose filters (16), and immunostained with the affinity-purified rabbit anti-human P4.2 IgG and a goat anti-rabbit IgG conjugated with alkaline phosphatase (Promega).

Generation of Cyanogen Bromide-Cleaved Fragments. Purified P4.2 (50 μg) prepared according to Korsgren and Cohen (7) was incubated with cyanogen bromide (17). The cleaved P4.2 mixture was concentrated in a Speed Vac Concentrator (Savant), and the peptide fragments were separated by HPLC (Aquapore C03-GU, 30 × 4.6 mm; Brownlee Lab) with a 0.1% trifluoroacetic acid/acetonitrile linear gradient (0–90% acetonitrile in 13 min) at a flow rate of 1 ml/min. The peptide peaks detected at 210 nm were collected, concentrated, and sequenced using an ABI 1470 gas phase protein sequenator (Applied Biosystems).

RNA Isolation and RNA Blot Analysis. Human reticulocyte RNA was prepared (8) from peripheral blood of an anemic individual with paroxysmal nocturnal hemoglobinuria. The RNA was electrophoresed on a 1% agarose/formaldehyde gel (18), transblotted onto nitrocellulose paper, and hybridized with ^{32}P-labeled probes generated by random-primer extension (Pharmacia LKB).

RESULTS

Isolation of P4.2 cDNA Clones. Immunoscreening of 5×10^5 recombinant phages with anti-human P4.2 antibody yielded 12 potentially positive clones, 6 of which contained large-sized inserts ranging from 1.2 to 1.8 kilobases (kb). The 1.8-kb insert of c.7 was found to cross-hybridize with those of c.4 (1.4 kb), c.8 (1.7 kb), and c.9 (1.2 kb), suggesting that these four clones contained overlapping nucleotide sequences.

Sequence analysis indicated that c.7 had a poly(A) tail and one long open reading frame (ORF) at the 5′ end. Restriction analyses and partial nucleotide sequences showed that these cross-hybridizing clones contained the 3′ portion of the cDNA with varying lengths from the 5′ end (Fig. 1). Since P4.2 has an apparent molecular mass of 72 kDa, its cDNA is expected to have a coding region of ≈2.2 kb. Hence, the longest c.7 of 1.8 kb was not long enough for the entire coding region.

5′-End Extension of cDNA. To obtain the missing 5′ end of the cDNA, we performed PCR 5′-end extension by using primers synthesized according to the sequences of c.7 insert and the λgt11 clones in the cDNA library. Sequencing of the four largest PCR-extended cDNA clones (including c.12 and c.16 in Fig. 1) showed that they had identical sequences, except that the largest one (c.12) had a 90-bp insert near the 5′ end of the coding region (Fig. 1). The combined length of c.7 and the 5′-extended cDNA is 2.4 kb (or 2.5 kb with the 90-bp insert). Sequence analyses showed that they contained ORFs of 2.1 and 2.2 kb, respectively, and thus were capable of encoding a protein of ≈72 kDa.

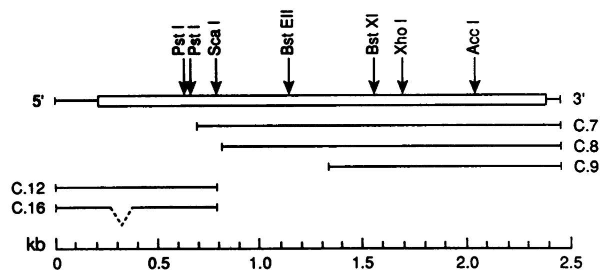

FIG. 1. Schematic diagram of human RBC P4.2 cDNAs. Horizontal open bar represents the coding region. Lines flanking it represent the 5′ and 3′ untranslated regions. Three clones isolated by immunoscreening (c.7, c.8, and c.9) and two clones obtained by PCR extension (c.12 and c.16) are shown below. The dashed V-line on c.16 indicates the absence of the 90-bp insert found in c.12.

Expression of Fusion Proteins and Amino Acid Sequence Analysis of Cyanogen Bromide Fragments. In addition to the coding capacities of the cDNAs, two other lines of evidence support the identity of these cDNAs as P4.2. An ≈175-kDa β-galactosidase fusion protein encoded by recombinant phage c.7 was detected by anti-P4.2 antibody (data not shown). Since β-galactosidase contributes 114 kDa to the fusion protein, c.7 insert encodes a peptide of ≈61 kDa. The combined protein size of c.7 and c.16 (15 kDa) or c.12 (18 kDa) is close to the apparent molecular mass of P4.2. The most convincing evidence for the authenticity of the cDNAs was the complete match of 46 amino acids from three independent cyanogen bromide peptides of P4.2 with the amino acid sequence deduced from the cDNA (Fig. 2, boxes).

Sequence Analysis of the P4.2 cDNAs. The complete nucleotide sequence of P4.2 cDNA and the deduced amino acid sequence of P4.2 are shown in Fig. 2. The cDNA has a 227-nt untranslated region upstream from the putative ATG start codon. The nucleotide sequence C<u>A</u>ACC ATG G around this initiation site is similar to the consensus sequence for initiation found in higher eukaryotes (19), except that the second nt in the P4.2 cDNAs is A rather than C. This ATG initiation site is followed by an ORF through the c.7 cDNA. There is another ATG at nt −179 to −177, but it is followed by an in-frame termination codon 19 nt downstream. The presence or absence of the 90-nt insert (underlined in Fig. 2) gives rise to two P4.2 cDNA sequences. The 2.4-kb cDNA contains 2382 nt with an ORF of 691 amino acids, predicting an ≈77-kDa protein; the 2.5-kb cDNA contains 2472 nt with an ORF of 721 amino acids, predicting an ≈80-kDa protein. The cDNA ends in a poly(A) tail, and the 3′ untranslated region is relatively short, containing only 82 nt. There is no polyadenylylation signal sequence AATAAA, but a sequence AA<u>TCT</u>AAA is located at nt 2204–2211.

RNA Blot Analysis of Human Reticulocytes. Northern blot analysis using c.7 insert as a probe detected a 2.4-kb RNA species in human reticulocytes (Fig. 3). This result indicates that the cDNAs obtained in this study (2382 and 2472 bp) are apparently the full-length cDNAs for P4.2. Overexposure of the blot showed two additional minor bands of 3.9 and 1.7 kb.

Structural Analysis of P4.2. The amino acid sequence derived from the 2.5-kb cDNA contains ≈43% nonpolar, ≈35% polar, ≈10% acidic, and ≈12% basic amino acid residues. The most abundant amino acids are leucine (82 residues) and alanine (60 residues). There are 49 serine and 43 threonine residues (potential sites for O-glycosylation), representing 13% of the total residues. There are 16 cysteine residues, 6 potential N-glycosylation sites (Asn-Xaa-Ser/Thr) at Asn-103, -420, -447, -529, -604, and -705, 1 potential cAMP-dependent phosphorylation site (basic-basic-Xaa-Ser) at Ser-278 (20), and 9 potential protein kinase C phosphorylation sites (Ser/Thr-Xaa-Arg/Lys) at Ser-7, -57, -58, -154, -222, -449, -455, and -666, and Thr-287 (21). There is one Arg-Gly-Asp sequence at 518–520. Secondary structure analysis using the Chou and Fasman method (22) predicted that P4.2 contains ≈33% β-sheet, ≈24% α-helix, and ≈45% reverse turns.

Hydropathy analysis of the deduced amino acid sequence using the algorithm and hydropathy values of Kyte and Doolittle (23) revealed a major hydrophobic domain (residues 298–322; Fig. 4, b). This hydrophobic region was predicted to be mainly a β-sheet structure with a possible turn. There is a strongly hydrophilic region (residues 438–495; Fig. 4, c). Toward the C terminus of this region, there is a highly charged segment predicted to be an α-helix (residues 470–492; underlined in Fig. 5) and containing a large number of both positively and negatively charged residues, especially glutamic acid.

Homology searches of GenBank 59 (released March 1989) and NBRF-PIR Protein Sequence Database (release 19,

```
                     GA  ACT GGT CCA GGA GAT  ACC AGC AGA GAG GGA  GTA GGA GAG AAG AAA  CAT GTC AGG GTG CTC  ACA GGA GTA GTG GGG  GGA GGT TTT GCT ATT  TCC AGA TTC TTA AGC -121
CAA CAA AAG TGC CTT  CAT ATT TTC TGT CTG  GAA GAC AGA AAG CCC  AGA AGG AGC CCA GAA  GCA ACA GTT TGA GAG  AGG CGC TTT CTG CGG  CCA AGT GGA TAA GAG  GAG CGG CCT GCA ACC -1
                                     10                                        20                                        30                                        40
Met Gly Gln Gly Glu  Pro Ser Gln Arg Ser  Thr Gly Leu Ala Gly  Leu Tyr Ala Ala Pro  Ala Ala Ser Pro Val  Phe Ile Lys Gly Ser  Gly Met Asp Ala Leu  Gly Ile Lys Ser Cys
ATG GGA CAG GGT GAG  CCA AGC CAG CGC TCG  ACA GGG CTT GCT GGA  CTG TAT GCA GCC CCC  GCA GCA TCA CCT GTT  TTC ATT AAA GGA AGT  GGG ATG GAT GCC CTG  GGT ATC AAG AGC TGT 120
                                     50                                        60                                        70                                        80
Asp Phe Gln Ala Ala  Arg Asn Asn Glu Glu  His His Thr Lys Ala  Leu Ser Ser Arg Arg  Leu Phe Val Arg Arg  Gly Gln Pro Phe Thr  Ile Ile Leu Tyr Phe  Arg Ala Pro Val Arg
GAC TTT CAG GCA GCA  AGA AAC AAT GAG GAG  CAC CAC ACC AAG GCC  CTC AGC TCC CGG CGC  CTC TTT GTG AGG AGG  GGG CAG CCC TTC ACC  ATC ATC CTG TAC TTC  CGC GCT CCA GTC CGT 240
                                     90                                       100                                       110                                       120
Ala Phe Leu Pro Ala  Leu Lys Lys Val Ala  Leu Thr Ala Gln Thr  Gly Glu Gln Pro Ser  Lys Ile Asn Arg Thr  Gln Ala Thr Phe Pro  Ile Ser Ser Leu Gly  Asp Arg Lys Trp Trp
GCA TTT CTG CCT GCC  CTG AAG AAG GTG GCC  CTC ACT GCA CAA ACT  GGA GAG CAG CCT TCC  AAG ATC AAC AGG ACC  CAA GCC ACA TTC CCA  ATT TCC AGT CTG GGG  GAC CGA AAG TGG TGG 360
                                    130                                       140                                       150                                       160
Ser Ala Val Val Glu  Glu Arg Asp Ala Gln  Ser Trp Thr Ile Ser  Val Thr Thr Pro Ala  Asp Ala Val Ile Gly  His Tyr Ser Leu Leu  Leu Gln Val Ser Gly  Arg Lys Gln Leu Leu
AGT GCA GTG GTG GAG  GAG AGA GAT GCC CAG  TCC TGG ACC ATC TCT  GTG ACC ACA CCT GCA  GAC GCT GTC ATT GGC  CAC TAC TCG CTT CTG  CTG CAG GTC TCA GGC  AGG AAG CAA CTC CTC 480
                                    170                                       180                                       190                                       200
Leu Gly Gln Phe Thr  Leu Leu Phe Asn Pro  Trp Asn Arg Glu Asp  Ala Val Phe Leu Lys  Asn Glu Ala Gln Arg [Met Glu Tyr Leu Leu  Asn Gln Asn Gly Leu  Ile Tyr Leu Gly] Thr
TTG GGT CAG TTC ACA  CTG CTT TTT AAC CCC  TGG AAT AGA GAG GAT  GCT GTG TTC CTG AAG  AAT GAG GCT CAG CGC [ATG GAG TAC TTG TTG  AAC CAG AAT GGT CTC  ATC TAC CTG GGT] ACA 600
                                    210                                       220                                       230                                       240
Ala Asp Cys Ile Gln  Ala Glu Ser Trp Asp  Phe Gly Gln Phe Glu  Gly Asp Val Ile Asp  Leu Ser Leu Arg Leu  Leu Ser Lys Asp Lys  Gln Val Glu Lys Trp  Ser Gln Pro Val His
GCT GAC TGC ATC CAG  GCA GAG TCC TGG GAC  TTT GGC CAG TTC GAG  GGG GAT GTC ATT GAC  CTC AGC CTG CGC TTG  CTG AGC AAG GAC AAG  CAG GTA GAG AAG TGG  AGC CAG CCG GTG CAC 720
                                    250                                       260                                       270                                       280
Val Ala Arg Val Leu  Gly Ala Leu Leu His  Phe Leu Lys Glu Gln  Arg Val Leu Pro Thr  Pro Gln Thr Gln Ala  Thr Gln Glu Gly Ala  Leu Leu Asn Lys Arg  Arg Gly Ser Val Pro
GTG GCC CGT GTG TTG  GGT GCC TTG CTG CAT  TTT CTC AAG GAG CAG  AGG GTC CTG CCC ACC  CCG CAG ACC CAG GCC  ACC CAG GAA GGG GCC  TTG CTG AAC AAG CGC  CGG GGC AGC GTG CCC 840
                                    290                                       300                                       310                                       320
Ile Leu Arg Gln Trp  Leu Thr Gly Arg Gly  Arg Pro Val Tyr Asp  Gly Gln Ala Trp Val  Leu Ala Ala Val Ala  Cys Thr Val Leu Arg  Cys Leu Gly Ile Pro  Ala Arg Val Val Thr
ATC CTG CGG CAG TGG  CTC ACC GGC CGA GGC  CGA CCT GTG TAT GAT  GGC CAG GCC TGG GTG  TTG GCT GCT GTT GCT  TGC ACA GTG CTG CGA  TGC CTG GGA ATC CCT  GCC CGC GTG GTG ACC 960
                                    330                                       340                                       350                                       360
Thr Phe Ala Ser Ala  Gln Gly Thr Gly Gly  Arg Leu Leu Ile Asp  Glu Tyr Tyr Asn Glu  Glu Gly Leu Gln Asn  Gly Glu Gly Gln Arg  Gly Arg Ile Trp Ile  Phe Gln Thr Ser Thr
ACG TTT GCC TCA GCA  CAG GGC ACC GGT GGG  CGT CTT CTC ATA GAT  GAA TAC TAT AAT GAG  GAG GGA CTT CAG AAC  GGA GAA GGC CAG AGA  GGC AGA ATC TGG ATC  TTC CAG ACT TCC ACA 1080
                                    370                                       380                                       390                                       400
Glu Cys Trp [Met Lys  Arg Gly Leu Pro Cys  Gln Gly Tyr Asp Gly  Trp Gln Ile Leu His] Pro Ser Ala Pro Asn  Gly Gly Gly Val Leu  Gly Ser Cys Asp Leu  Val Pro Val Arg Ala
GAG TGC TGG [ATG AAG  CGC GGC CTG CCT TGC  CAG GGT TAT GAT GGA  TGG CAG ATT CTG CAC] CCA AGT GCT CCT AAT  GGA GGT GGA GTC CTG  GGG TCC TGT GAT CTG  GTG CCG GTC AGA GCA 1200
                                    410                                       420                                       430                                       440
Val Lys Glu Gly Thr  Leu Gly Leu Thr Pro  Ala Val Ser Asp Leu  Phe Ala Ala Ile Asn  Ala Ser Cys Val Val  Trp Lys Cys Cys Glu  Asp Gly Thr Leu Glu  Leu Thr Asp Ser Asn
GTC AAG GAG GGG ACG  CTG GGG CTG ACC CCA  GCA GTG TCA GAC CTT  TTT GCT GCC ATA AAT  GCC TCA TGT GTG GTC  TGG AAG TGC TGT GAG  GAT GGG ACA CTG GAG  TTG ACT GAC TCC AAC 1320
                                    450                                       460                                       470                                       480
Thr Lys Tyr Val Gly  Asn Asn Ile Ser Thr  Lys Gly Val Gly Ser  Asp Arg Cys Glu Asp  Ile Thr Gln Asn Tyr  Lys Tyr Pro Glu Gly  Ser Leu Gln Glu Lys  Glu Val Leu Glu Arg
ACA AAG TAT GTT GGC  AAC AAC ATC AGC ACC  AAG GGT GTG GGC AGT  GAC CGC TGC GAG GAC  ATC ACT CAG AAC TAC  AAG TAT CCT GAA GGG  TCT CTT CAG GAA AAA  GAG GTG CTG GAG AGA 1440
                                    490                                       500                                       510                                       520
Val Glu Lys Glu Lys  Met Glu Arg Glu Lys  Asp Asn Gly Ile Arg  Pro Pro Ser Leu Glu  Thr Ala Ser Pro Leu  Tyr Leu Leu Leu Lys  Ala Pro Ser Ser Leu  Pro Leu Arg Gly Asp
GTC GAG AAA GAG AAA  ATG GAA CGT GAG AAA  GAC AAC GGC ATC CGT  CCT CCC AGT CTC GAG  ACT GCC AGT CCT CTG  TAC CTG CTC TTG AAA  GCA CCC AGC TCC CTA  CCC CTG AGA GGG GAT 1560
                                    530                                       540                                       550                                       560
Ala Gln Ile Ser Val  Thr Leu Val Asn His  Ser Glu Gln Glu Lys  Ala Val Gln Leu Ala  Ile Gly Val Gln Ala  Val His Tyr Asn Gly  Val Leu Ala Ala Lys  Leu Trp Arg Lys Lys
GCC CAG ATC TCA GTG  ACG CTG GTT AAT CAC  AGT GAG CAG GAG AAG  GCA GTG CAG CTG GCA  ATT GGG GTC CAG GCT  GTA CAC TAC AAC GGT  GTC CTT GCT GCC AAG  CTC TGG AGG AAG AAG 1680
                                    570                                       580                                       590                                       600
Leu His Leu Thr Leu  Ser Ala Asn Leu Glu  Lys Ile Ile Thr Ile  Gly Leu Phe Phe Ser  Asn Phe Glu Arg Asn  Pro Pro Glu Asn Thr  Phe Leu Arg Leu Thr  Ala Met Ala Thr His
CTG CAC CTC ACG CTC  AGT GCC AAC CTG GAA  AAG ATA ATA ACC ATC  GGC CTG TTC TTC TCC  AAT TTT GAG CGA AAC  CCA CCC GAG AAC ACC  TTC CTT AGA CTC ACC  GCC ATG GCA ACA CAC 1800
                                    610                                       620                                       630                                       640
Ser Glu Ser Asn Leu  Ser Cys Phe Ala Gln  Glu Asp Ile Ala Ile  Cys Arg Pro His Leu  Ala Ile Lys [Met Pro  Glu Lys Ala Glu Gln  Tyr Gln Pro Leu Thr  Ala Ser Val] Ser Leu
TCT GAA TCC AAC CTT  AGC TGC TTT GCT CAG  GAA GAC ATT GCC ATT  TGT AGA CCA CAC CTT  GCC ATC AAG [ATG CCA  GAG AAA GCA GAG CAG  TAT CAA CCC CTC ACA  GCC TCA GTC] AGC CTC 1920
                                    650                                       660                                       670                                       680
Gln Asn Ser Leu Asp  Ala Pro Met Glu Asp  Cys Val Ile Ser Ile  Leu Gly Arg Gly Leu  Ile His Arg Glu Arg  Ser Tyr Arg Phe Arg  Ser Val Trp Pro Glu  Asn Thr Met Cys Ala
CAG AAC TCC CTA GAT  GCC CCC ATG GAG GAC  TGT GTG ATC TCC ATC  CTG GGA AGG GGG CTC  ATT CAC AGA GAG AGG  AGC TAC AGA TTC CGT  TCA GTG TGG CCT GAA  AAC ACC ATG TGT GCC 2040
                                    690                                       700                                       710                                       720
Lys Phe Gln Phe Thr  Pro Thr His Val Gly  Leu Gln Arg Leu Thr  Val Glu Val Asp Cys  Asn Met Phe Gln Asn  Leu Thr Asn Tyr Lys  Ser Val Thr Val Val  Ala Pro Glu Leu Ser
AAG TTC CAG TTC ACG  CCA ACA CAT GTG GGG  CTC CAG AGA CTC ACT  GTG GAA GTG GAC TGC  AAC ATG TTC CAG AAC  CTA ACC AAC TAT AAA  AGC GTC ACC GTG GTA  GCC CCT GAA CTA TCA 2160

Ala
GCT TAA ACT TCC AGC TCT ATC  ACC ACT CTC CTG CCA  ACC CTT GTT CTA CAA  TCT AAA CCA  AAC ATG TGC TAG GAA GAG AAA  AAA A
```

FIG. 2. Composite nucleotide sequence of P4.2 cDNA and deduced amino acid sequence. The first nucleotide of the putative ATG initiation codon is designated as nucleotide position 1. The underlined (without arrow) sequence represents the 90-bp insert in the larger cDNA. Boxed residues are those matched with the amino acid sequences obtained from three independent cyanogen bromide-cleaved peptides of purified P4.2. Positions where the primer sequences were derived for PCR extension are underlined with double arrowheads (p1) and a single arrowhead (p2).

updated December 1988) revealed that the amino acid sequence of human RBC P4.2 has significant homology with subunit a of human placental factor XIII ($XIII_a$, plasma transglutaminase; refs. 24, 26, 27) and also with the guinea pig liver transglutaminase (LTG; ref. 25) (Fig. 5). While the two transglutaminases are themselves 35.3% identical, P4.2 has 24.6% and 33.5% identity with $XIII_a$ and LTG, respectively. Fig. 5 shows the alignment of these three proteins by the method of Feng and Doolittle (28). In this alignment, a total of 122 residues are identical in all three proteins (Fig. 5, asterisks). The strongest homology is found in residues 293–325 of P4.2 (Fig. 5, large box), in which 20 of the 33 residues, or 60.6%, are identical for all three proteins. This highly conserved region contains the catalytic thiol sites of $XIII_a$ and LTG where Cys-Trp is required for transglutaminase activity (24, 26). P4.2, however, despite the high homology around the active site, has an alanine (residue 298 of P4.2; Fig. 5, arrow) instead of cysteine at the corresponding position.

The hydropathy plots for $XIII_a$, LTG, and P4.2 show similarities (Fig. 4), especially with regard to the hydrophobic region labeled b (residues 298–322 in P4.2), which overlaps with the 60.6% identity area (residues 293–325 in P4.2) in the larger box of Fig. 5. This region begins with the critical cysteine residue in $XIII_a$ and LTG and the corresponding alanine in P4.2. In the region marked c in Fig. 4, P4.2 has a strongly hydrophilic region; it is less hydrophilic in LTG and least in $XIII_a$. Near the center of this hydrophilic region in P4.2 lies a potential Ca^{2+} binding site (Fig. 5, small box), as identified by its alignment and homology with $XIII_a$ and other Ca^{2+} binding proteins (24).

FIG. 3. Northern blot analysis of human reticulocyte RNA. A major 2.4-kb hybridizing RNA was detected with c.7 probe (see Fig. 1). The other two arrows indicate the minor bands (3.9 and 1.7 kb) that appeared after overexposure. The standards used for size estimation were 18S rRNA (≈1.8 kb) and 28S rRNA (≈4.8 kb). The line at the top of the gel represents the origin.

DISCUSSION

In this communication, we report the cloning and sequencing of cDNA encoding the human RBC membrane P4.2. Two P4.2 cDNA sequences were obtained: one contains 2382 nt with an ORF of 691 residues; the other, with a 90-nt insert, contains 2472 nt with an ORF of 721 residues. These predictions are in good agreement with the apparent molecular mass of P4.2 (≈72 kDa). The finding of the two P4.2 cDNAs is consistent with the immunostaining pattern of P4.2 on SDS/

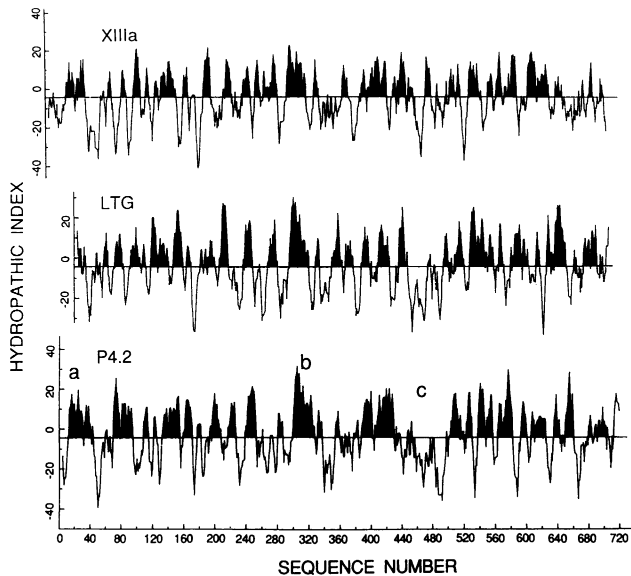

FIG. 4. Hydropathy plot of the deduced amino acid sequence of human P4.2 (with regions marked a, b, and c) and its comparison with subunit a of human factor XIII and guinea pig liver transglutaminase LTG. The hydropathic index was obtained from windows of seven amino acids. The three plots are aligned according to the highly conserved hydrophobic region (designated b: residues 298–322 in P4.2), which contains the transglutaminase active site in $XIII_a$ (24) and LTG (25). The catalytic cysteine of the active sites of the transglutaminases and the corresponding alanine in P4.2 (see arrow in large box in Fig. 5) are all located at the point of transition from hydrophilic to hydrophobic regions—i.e., the beginning of area b. In the P4.2 panel, area a (residues 4–33) shows the hydrophobic characteristics of a 30-amino acid insert in the 2.5-kb cDNA, and area c (residues 438–495) shows the strongly hydrophilic region of P4.2.

PAGE, which showed a diffuse band slightly higher than the major 72-kDa band (data not shown).

PCR-extended sequences contain 44 nt (including the 17-nt primer) overlapping with the 5′ end of c.7 obtained directly from the cDNA library. Furthermore, both the PCR-extended sequences (after removal of the nucleotides overlapping with c.7) and c.7 insert hybridized strongly with a P4.2 partial genomic DNA clone (unpublished observation). All of the above show that the PCR products were part of the P4.2 gene. The c.12, however, does not have the first 7 nt (ACAAACT) at the 5′ end of c.7. These 7 nt may represent the end of another insertion/deletion sequence that was not amplified and subcloned during PCR extension. Interestingly, an ACAAACT sequence is found further upstream at nt 279–285.

The mechanism by which the two P4.2 isoforms arise is unknown. Alternative splicing is an attractive possibility, especially in light of recent findings that isoforms of protein 4.1 mRNAs are generated by such a mechanism (9–11). The junction sequence around the 90-bp insert are G/G and T/G, which have been reported as junction nucleotides between exons (29, 30). This 30-amino acid insert has the characteristics of signal peptides containing a stretch of hydrophobic residues and shows homology with the internal sequences of a group of tyrosine kinase-related transforming proteins—e.g., c-src (31). This 30-amino acid insert may represent an imported exon. The expression and possible function of this insert warrants further study. RNA blot analysis indicates that the cDNAs obtained in this study represent the full-length message for P4.2 in reticulocytes, although the 90-nt difference of these two isoforms cannot be resolved. Whether the two minor hybridizing RNA species (3.9 and 1.7 kb) represent additional isoforms or messages of related proteins in reticulocytes needs to be investigated.

P4.2 binds to the cytoplasmic pole of band 3 in membranes (32, 33) and to ankyrin and protein 4.1 in solution (33). The availability of the P4.2 cDNAs has allowed us to conduct preliminary studies on the functional domains of P4.2, including those involved in its binding to other proteins. P4.2 contains 43% hydrophobic amino acid residues and shows at

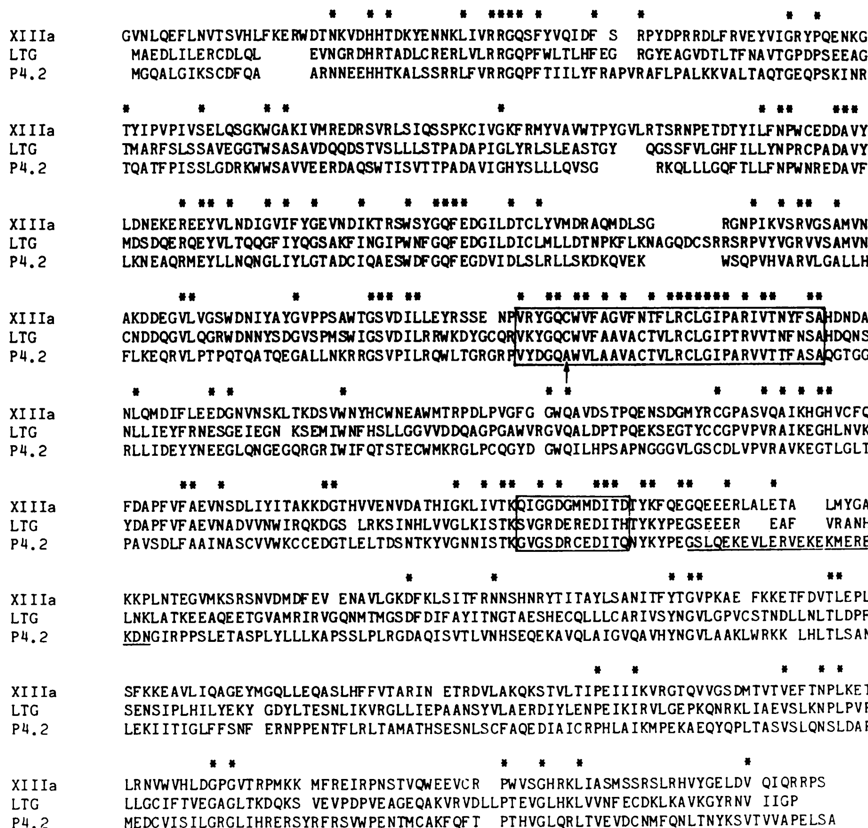

FIG. 5. The alignment of human P4.2 with $XIII_a$ (694 amino acids) and LTG (689 amino acids). Asterisks indicate residues that are identical in all three proteins. The alignment necessitates removal of the 30-amino acid insertion from the long isoform of P4.2—i.e., the shorter isoform (691 amino acids) is used in this plot. Gaps are allowed for maximizing identity. The region with the greatest identity is enclosed in the larger box (residues 293–325 of P4.2), which contains the highly conserved active site with Cys-Trp (CW) in $XIII_a$ and LTG (24). Arrow indicates the presence of alanine in P4.2 instead of cysteine at this site. Smaller boxed area (residues 452–463 of P4.2) is the potential Ca^{2+} binding site of $XIII_a$. Underlined sequence (residues 470–492 of P4.2) is the highly charged region in P4.2 with a predicted α-helical structure flanked by nearby proline residues. Amino acids are designated by the single-letter code.

least one domain of high hydrophobicity (Fig. 4, b). The hydrophobic region(s) might interact with membrane lipids or allow P4.2 to fold within itself. Since the hydrophobic domain labeled b is highly conserved in P4.2 and the two transglutaminases, it may be important in forming the active site itself and/or positioning the sites in cells.

It is interesting that P4.2 contains one potential cAMP-dependent phosphorylation site, since Suzuki *et al.* (34) have previously reported that P4.2 was phosphorylated by a cAMP-dependent protein kinase and that phosphorylation was stimulated by heavy metal ions.

P4.2 has significant homology with $XIII_a$ (24, 26, 27) and LTG (25), especially around their active sites. $XIII_a$, the final component in the coagulation pathway, plays an important role in the stabilization of fibrin clots by covalently crosslinking fibrin monomers through γ-glutamyl-ε-lysine bridges and by preventing proteolysis (35). Liver transglutaminase has some of the activity associated with the plasma membrane and may be responsible for forming covalently crosslinked matrices of proteins at sites of cell-to-cell contact (36). P4.2, however, has alanine instead of the cysteine indispensible for transglutaminase activity (37) in the active site area. It is possible that P4.2 may use this site to bind other RBC membrane proteins without forming covalent crosslinks. P4.2, along with protein 4.1, has been proposed to be one of the last membrane proteins synthesized during RBC maturation (38). P4.2 may contribute to the stabilization of the membrane skeleton through its binding with membrane proteins and thus protects them from being degraded (e.g., by proteases) or crosslinked (e.g., by cytoplasmic transglutaminase). Many proteins that are labile in the cytosol become resistant to degradation once assembled into the skeletal network (39).

The phylogenetic history of the three proteins can be inferred from their relative similarities. The human P4.2 is more similar to LTG than $XIII_a$, but it appears to have undergone a faster rate of change, suggesting that P4.2 is an offshoot of a tissue enzyme.

Our finding of P4.2 cDNAs in the human reticulocyte cDNA library indicates that circulating reticulocytes retain intact P4.2 mRNAs. Immunoreactive analogs of P4.2 are also present in nonerythroid cells and tissues, including platelets, brain, and kidney (40, 41). These results raise the possibility that P4.2, like protein 4.1, may be a ubiquitous component of cell membranes, although its function in other cells may differ from that in RBCs. The availability of the cDNA for P4.2 should aid considerably in the study of structure–function relationships of this protein, including investigations on the mechanisms responsible for P4.2 deficiency in human patients and the expression of P4.2 in different tissues and during differentiation.

Note. Korsgren *et al.* (42) have also obtained a cDNA sequence of human RBC membrane P4.2. Their nucleotide sequence, except for the following, is identical to our shorter isoform: (*i*) Absence of the first 39 nucleotides at the 5′ end; (*ii*) C instead of A at −188 and absence of T at −37; (*iii*) four synonymous differences: G instead of A at 420, T vs. G at 531, and C vs. G at both 1137 and 1215; (*iv*) two nonsynonymous differences: G vs. C at both 1138 and 1216; (*v*) absence of AG at 1094–1095 and presence of extra CC after 1109, causing a frameshift in the intervening nucleotides. *iv* and *v* result in seven amino acid differences.

We thank Drs. John G. Conboy, Y. W. Kan, and Narla Mohandas for their generous gift of the cDNA library, Drs. Al Smith and John Gardner for amino acid sequence analysis of P4.2 peptides, Dr. Thomas Shenk for his laboratory facilities, Dr. Michael G. Rosenfeld for his valuable advice and help throughout this study, Dr. John Trombold for blood samples for RNA preparations, and Dr. Russell F. Doolittle for his expert advice and help in alignment and hydropathy analysis of proteins. We appreciate the technical expertise of Sylvia Musto, June Wang, Eugene Leung, and Gerard Norwich. This work was supported by Research Grants HL19454, HL21016, HL33084, HL38655, and HL44147 from the National Heart, Lung and Blood Institute.

1. Bennett, V. (1985) *Annu. Rev. Biochem.* **54,** 273–304.
2. Becker, P. S. & Benz, E. J. (1990) in *Molecular Biology of the Cardiovascular System*, ed. Chien, S. (Lea & Febiger, Philadelphia), in press.
3. Curtis, P. J., Palumbo, A., Ming, J., Fraser, P., Cioe, L., Meo, P., Shane, S. & Rovera, G. (1985) *Gene* **36,** 357–362.
4. Kopito, R. R. & Lodish, H. F. (1985) *Nature (London)* **316,** 234–238.
5. Conboy, J., Kan, Y. W., Shohet, S. B. & Mohandas, N. (1986) *Proc. Natl. Acad. Sci. USA* **83,** 9512–9516.
6. McMahon, A. P., Giebelhaus, D. H., Champion, J. E., Bailes, J. A., Lacey, S., Carritt, B., Henchman, S. K. & Moon, R. T. (1987) *Differentiation* **34,** 68–78.
7. Korsgren, C. & Cohen, C. M. (1986) *J. Biol. Chem.* **261,** 5536–5543.
8. Rybicki, A. C., Heath, R., Wolf, J. L., Lubin, B. & Schwartz, R. S. (1988) *J. Clin. Invest.* **81,** 893–901.
9. Ngai, J., Stack, J. H., Moon, R. T. & Lazarides, E. (1987) *Proc. Natl. Acad. Sci. USA* **84,** 4432–4436.
10. Tang, T., Leto, T. L., Correas, I., Alonso, M. A., Marchesi, V. T. & Benz, E. J. (1988) *Proc. Natl. Acad. Sci. USA* **85,** 3713–3717.
11. Conboy, J. G., Chan, J., Mohandas, N. & Kan, Y. W. (1988) *Proc. Natl. Acad. Sci. USA* **85,** 9062–9065.
12. Huynh, T. V., Young, R. A. & Davis, R. W. (1985) in *DNA Cloning: A Practical Approach*, ed. Glover, D. M. (IRL, Oxford, U.K.), Vol. 1, pp. 49–78.
13. Maniatis, T., Fritsch, E. F. & Sambrook, J. (1982) *Molecular Cloning: A Laboratory Manual* (Cold Spring Harbor Lab., Cold Spring Harbor, NY), pp. 135–139.
14. Sanger, F., Nicklen, S. & Coulson, A. R. (1977) *Proc. Natl. Acad. Sci. USA* **74,** 5463–5467.
15. Laemmli, U. K. (1970) *Nature (London)* **227,** 680–685.
16. Towbin, H., Staehelin, T. & Gordon, J. (1979) *Proc. Natl. Acad. Sci. USA* **76,** 4350–4354.
17. Drickamer, L. K. (1977) *J. Biol. Chem.* **252,** 6909–6917.
18. Ausubel, F. M., Brent, R., Kingston, R. G., Moore, D. D., Smith, J. A., Seidman, J. G. & Struhl, K. (1987) *Current Protocols in Molecular Biology* (Wiley, New York), p. 4.1.2.
19. Kozak, M. (1987) *J. Mol. Biol.* **196,** 947–950.
20. Krebs, E. G. & Beavo, J. A. (1979) *Annu. Rev. Biochem.* **48,** 923–959.
21. Woodget, J. R., Gould, K. L. & Hunter, T. (1986) *Eur. J. Biochem.* **161,** 177–184.
22. Chou, P. Y. & Fasman, G. D. (1974) *Biochemistry* **13,** 222–245.
23. Kyte, J. & Doolittle, R. F. (1982) *J. Mol. Biol.* **157,** 105–132.
24. Takahashi, N., Takahashi, Y. & Putnam, F. W. (1986) *Proc. Natl. Acad. Sci. USA* **83,** 8019–8023.
25. Ikura, K., Nasu, C., Yokota, H., Tsuchiya, Y., Sasaki, R. & Chiba, H. (1988) *Biochemistry* **27,** 2898–2905.
26. Ichinose, A., Hendrickson, L. E., Fujikawa, K. & Davie, E. W. (1986) *Biochemistry* **25,** 6900–6906.
27. Grundmann, U., Amann, E., Zettlmeissl, G. & Kupper, H. A. (1986) *Proc. Natl. Acad. Sci. USA* **83,** 8024–8028.
28. Feng, D.-F. & Doolittle, R. F. (1987) *J. Mol. Evol.* **25,** 251–260.
29. Yamada, Y., Avvedimento, V. E., Mudryi, M., Ohkubo, H., Vogeli, G., Meher, I., Pastan, I. & deCrombrugghe, B. (1980) *Cell* **22,** 887–892.
30. Mount, S. M. (1982) *Nucleic Acids Res.* **10,** 459–472.
31. Hanks, S. K., Quinn, A. M. & Hunter, T. (1988) *Science* **241,** 42–52.
32. Steck, T. L. (1974) *J. Cell Biol.* **62,** 1–19.
33. Korsgren, C. & Cohen, C. M. (1988) *J. Biol. Chem.* **263,** 10212–10216.
34. Suzuki, K., Ikebuchi, H. & Terao, T. (1985) *J. Biol. Chem.* **260,** 4526–4530.
35. Lorand, L., Downey, J., Gotoh, T., Jacobson, A. & Tokura, S. (1968) *Biochem. Biophys. Res. Commun.* **31,** 222–230.
36. Slife, C. W., Dorsett, M. D. & Tilotson, M. L. (1986) *J. Biol. Chem.* **261,** 3451–3456.
37. Folk, J. E. (1980) *Annu. Rev. Biochem.* **49,** 517–531.
38. Chang, H., Langer, P. J. & Lodish, H. F. (1976) *Proc. Natl. Acad. Sci. USA* **73,** 3206–3210.
39. Moon, R. T. & Lazarides, E. (1984) *J. Cell Biol.* **98,** 1899–1904.
40. Schwartz, R. S., Rybicki, A. C., Heath, R., Shew, R. & Lubin, B. (1987) *Blood* **70,** Suppl. 1, 42a (abstr.).
41. Friedrichs, B., Koob, R., Kraemer, D. & Drenckhahn, D. (1989) *Eur. J. Cell Biol.* **48,** 121–127.
42. Korsgren, C., Lawler, J., Lambert, S., Speicher, D. & Cohen, C. M. (1990) *Proc. Natl. Acad. Sci. USA* **87,** 613–617.