Cloning and characterization of two cDNAs coding for human von Willebrand factor

(cDNA cloning/DNA sequence analysis)

J. Evan Sadler^{*†‡}, Beverly B. Shelton-Inloes^{*†‡}, James M. Sorace^{*§}, John M. Harlan[¶], Koiti Titani^{||}, and Earl W. Davie^{||}

*Howard Hughes Medical Institute Laboratory, Departments of †Medicine, ‡Biochemistry, and §Laboratory Medicine, Washington University School of Medicine, St. Louis, MO 63110; and Departments of ^{||}Biochemistry and ¶Medicine, University of Washington, Seattle, WA 98195

Contributed by Earl W. Davie, May 29, 1985

A cDNA library was prepared in Agt11 ABSTRACT bacteriophage from poly(A)⁺ RNA isolated from primary cultures of endothelial cells from human umbilical vein. Approximately 2.5 million independent recombinants were screened and 2 of those were found to synthesize a fusion protein with β -galactosidase that reacted with rabbit antibody against human von Willebrand factor. Comparison of the amino acid sequence translated from the cDNA insert of the two clones with the amino acid sequence determined by Edman degradation of the protein established that both phage isolates code for von Willebrand factor. The first clone (λ HvWF1) contained an insert of 404 nucleotides that corresponded to amino acid residues 1-110 in the mature protein circulating in blood, in addition to a portion (24 amino acids) of a prepro leader sequence. The second cDNA clone (\lambda HvWF3) contained an insert of 4.9 kilobases that coded for the carboxyl-terminal 1525 amino acids of von Willibrand factor, a stop codon of TGA, 134 nucleotides of 3' noncoding sequence, and a poly(A) tail of 150 nucleotides. The two clones together code for >80% of the molecule circulating in blood. The same carboxylterminal lysine residue was identified in the mature protein as well as in the cDNA, indicating that all of the proteolytic processing that occurs during the biosynthesis and assembly of von Willebrand factor is associated with the amino-terminal portion of the precursor protein. The amino acid sequence of von Willebrand factor indicates the presence of two different internal gene duplications and one triplication. These repetitive amino acid sequences account for about one-half of the amino acids present in the mature protein. The tetrapeptide sequence of Arg-Gly-Asp-Ser, which mediates the cell attachment and platelet binding activity of fibronectin, was also identified in the carboxyl-terminal portion of von Willebrand factor.

von Willebrand factor is a multimeric plasma glycoprotein that consists of subunits $(M_r, 260,000)$ that are held together by disulfide bonds. It circulates in blood as multimers that range in size from dimers of ≈500,000 to multimers of >10,000,000. von Willebrand factor is synthesized by endothelial cells throughout the body (1, 2) and also by megakaryocytes (3). It participates in the initial reactions of hemostasis by forming a bridge between platelets and the damaged vascular subendothelium, and this leads to platelet plug formation. Specific receptors for von Willebrand factor have been identified on the platelet membrane as well as the subendothelium. With platelets, its binding is associated with platelet glycoprotein IB (4-6) and, under some circumstances in vitro, to platelet glycoprotein IIB/IIIA (7, 8). The principal receptor for von Willebrand factor is probably collagen in the subendothelial connective tissue (9-11). von

Willebrand factor also forms a complex with factor VIII (antihemophilic factor), and this interaction is necessary for the survival of the coagulant protein *in vivo* (12). The function of von Willebrand factor in platelet plug formation appears to be dependent on the assembly of the protein subunits into large multimers. Accordingly, the decreased biological activity of the plasma protein may be due to abnormalities of polymerization or decreased levels of the protein in blood. Both types of defect have been described in von Willebrand's disease.

Small amounts of highly purified von Willebrand factor have been available for nearly 15 years, but the biochemical characterization of the protein has been hampered by its large size and complexity. Recently, several posttranslational modifications have been described during the biosynthesis of von Willebrand factor, including proteolytic processing (13, 14), glycosylation (13, 15, 16), and sulfation (17). Formation of disulfide bonds leads to the generation of the small and large multimers. Some relationships between protein domains, oligomeric structure, and platelet-related functions have been described (18–20). Many variants of von Willebrand's disease have been identified, and these have been classified on the basis of the structural properties of the mutant proteins as well as *in vitro* tests (8, 21, 22).

As a step toward understanding the structure-function relationships, gene organization, biosynthetic regulatory mechanisms, and evolution of von Willebrand factor, cDNA clones coding for von Willebrand factor have been isolated from a human endothelial cell cDNA library in our laboratory (23) as well as by others (24, 25). In this report, we describe the DNA sequence of two cDNA inserts that together code for >80% of the mature protein present in plasma, in addition to 24 amino acids from an amino-terminal leader peptide.

MATERIALS AND METHODS

Restriction endonucleases, nuclease BAL-31, T4 DNA ligase, T4 polynucleotide kinase, EcoRI methylase, and T4 DNA polymerase were purchased from Bethesda Research Laboratories or New England Biolabs. Human placental ribonuclease inhibitor was supplied by Bolton Biologicals (Richmond Heights, MO), and reverse transcriptase was purchased from Seikagaku America (St. Petersburg, FL). Oligo(dT)- cellulose, oligo(dT)₁₂₋₁₈, EcoRI linkers, ATP, dideoxynucleotide, and deoxynucleotide triphosphates were supplied by Pharmacia P-L Biochemicals. The Klenow fragment of Escherichia coli DNA polymerase was purchased from Bethesda Research Laboratories and Boehringer Mannheim. Calf intestine alkaline phosphatase and nuclease S1 were products of Boehringer Mannheim. Deoxyadenosine 5'-[[α -³⁵S]thio]triphosphate (dATP[α -³⁵S]) was purchased from Amersham. Na¹²⁵I and $[\gamma^{-32}P]ATP$ were purchased from New England Nuclear.

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.

For cDNA cloning, umbilical vein endothelial cells were cultured by standard methods (26, 27), and RNA was isolated by centrifugation through 5.7 M CsCl after lysis of the cells in 5 M guanidine thiocyanate (28). Poly(A)⁺ RNA was selected by chromatography on oligo(dT) cellulose. Doublestranded cDNA was synthesized by using reverse transcriptase for first-strand synthesis, and with both reverse transcriptase and Klenow fragment for second-strand synthesis. These reactions as well as methylation, addition of *Eco*RI linkers, ligation into the λ gt11 arms, and packaging *in vitro* were performed as described by Schwarzbauer *et al.* (29). From 10 μ g of poly(A)⁺ RNA, 0.42 μ g of size-selected double-stranded cDNA was prepared, yielding \approx 20 million independent recombinant phage.

Rabbit antiserum and affinity-purified antibodies to human von Willebrand factor were prepared as described by Canfield and Kisiel (30). The purified antibody was labeled with ¹²⁵I to a specific activity of 4 million cpm/ μ g (31) and was used to screen the library at a plating density of 50,000 recombinants per 150-mm plate (32). Positive clones were plaque-purified and DNA was prepared by a plate-lysis method (33, 34).

For DNA sequencing, the cDNA inserts from positive isolates were subcloned into plasmid pUC18 (35) and suitable fragments were further subcloned into M13 mp18 or M13 mp19 (35). Sequencing was performed by the dideoxy method (36) using dATP[α -³⁵S] and buffer-gradient gels (37). Controlled digestions with nuclease BAL-31 were used to generate templates providing overlapping sequences (38). When necessary, oligonucleotides (17-20 bases) were synthesized (Applied Biosystems, Foster City, CA; DNA synthesizer) and used as sequencing primers with appropriate templates. All sequences were determined at least once on each strand by the dideoxy method. The first 220 nucleotides in λ HvWF1 were also sequenced by the method of Maxam and Gilbert (39). DNA sequences were analyzed with the computer programs of CompuGene (St. Louis, MO) and the Protein Identification Resources (40).

RESULTS AND DISCUSSION

A human endothelial cell cDNA library prepared in λ gt11 was screened with affinity-purified, ¹²⁵I-labeled rabbit antibody to human von Willebrand factor. Among 2.5 million recombinants screened, 2 positive λ phage were identified and plaque purified. The DNA inserts in the two isolates were then subcloned into pUC18 for further characterization. The first isolate (λ HvWF1) contained a cDNA insert of 404 nucleotides flanked by *Eco*RI linkers (Fig. 1). The second isolate (λ HvWF3) contained an insert of 4.9 kilobases of DNA, but only one linker *Eco*RI site was reconstituted. The DNA in this phage was digested with *Sac* I, which led to the formation of two fragments that together spanned all but 280 nucleotides of the cDNA insert (Fig. 1). The remaining DNA fragment was sequenced from an *Acc* I/*Eco*RI subclone of λ HvWF3. The DNA and corresponding amino acid sequences of the DNA inserts in λ HvWF1 and λ HvWF3 are shown in Fig. 2.

 λ HvWF1 codes for the amino-terminal end of von Willebrand factor in addition to 24 amino acid residues that constitute a portion of a leader sequence. This conclusion was made possible by the fact that the amino-terminal sequence of the von Willebrand factor subunit has been established as Ser-Leu-Ser-Cys-Arg-Pro-Pro by automated Edman degradation of the intact protein and a cyanogen bromide fragment originating from the amino-terminal end of the protein (ref. 50; M. Chopek, personal communication). The amino acid residue occurring immediately before the amino-terminal serine was arginine. Since signal peptidase does not cleave Arg-Ser bonds, the partial leader sequence of 24 amino acids represents a portion of a prepro peptide (41).

The cDNA sequence in λ HvWF3 contained a long open reading frame extending from nucleotide 2 through nucleotide 4576. The last 11 amino acid residues in this reading frame corresponded exactly to the carboxyl-terminal amino acid sequence of von Willebrand factor as determined by automated Edman degradation of the CNBr peptide originating from the carboxyl-terminal end of the protein (M. Chopek, personal communication). Therefore, the DNA insert in λ HvWF3 corresponds to the last 1525 amino acids of the protein. The carboxyl-terminal lysine (AAG) was followed by a stop codon (TGA) and 134 nucleotides of 3' noncoding sequence. These data indicate that all of the proteolytic processing that occurred during the assembly of von Willebrand factor into multimers took place in the amino-terminal region of the precursor polypeptide. A polyadenylylation or processing signal of AATAAA was identified 25 nucleotides prior to the poly(A) tail of 150 nucleotides.

The cDNA inserts in λ HvWF1 and λ HvWF3 do not overlap. Between them, they code for 1635 amino acids present in the mature von Willebrand factor subunit present in plasma and account for all but three of the known CNBr fragments of the protein (unpublished results). The mature protein present in plasma consists of \approx 2000 amino acids (19). Thus, there is a gap of \approx 1 kilobase of DNA coding for \approx 350 amino acids that is not represented in either λ HvWF1 or λ HvWF3. Thus far, \approx 80% of the sequence of von Willebrand factor has been established by amino acid sequence analysis (unpublished results). The amino acid sequence predicted from the two cDNAs is in excellent agreement with these data except for histidine-7 in λ HvWF1, which was clearly a proline by amino acid sequence may



FIG. 1. Partial restriction maps of the cDNA inserts in λ HvWF1 and λ HvWF3 that code for von Willebrand factor. The cDNA inserts are shown with the 5' end of the coding strand at the left; thus, the conventional orientation of the left and right arms of the λ phage are reversed. Only those restriction sites used in subcloning and DNA sequencing are shown. BP, base pairs; KB, kilobase.

Ser Glu Ala Glu Phe Glu Val Leu Lys Ala Phe Val Val Asp Met Met Glu Arg Leu Arg Ile Ser Gln Lys Trp Val Arg Val Ala Val Val Glu Tyr His Asp Gly T ICC GAG GCT GAG TIT GAA GTG CTG AAG GCC TIT GTG GTG GAC ATG ATG GAG GGG CTG GGC ATC TCC CAG AAG TGG GTC GGC GTG GTG GTG GAG TAC CAC GAC GGC Ser His Ala Tyr lle Gly Leu Lys Asp Arg Lys Arg Pro Ser Glu Leu Arg Arg lle Ala Ser Gln Val Lys Tyr Ala Gly Ser Gln Val Ala Ser Thr Ser Glu Val Leu TCC CAC GCC TAC ATC GGG CTC AAG GAC CGG AAG CGA CCA TCA GAG CTG CGG CGC ATT GCC AGC CAG GTG AGG TAT GCG GGC CAC CAG GTG GCC TCC ACC AGC GAG GTC TTG 37 110 Lys Tyr Thr Leu Phe Gin 11e Phe Ser Lys 11e Asp Arg Pro Giu Ala Ser Arg 11e Ala Leu Leu Met Ala Ser Gin Giu Pro Gin Arg Met Ser Arg Asn Phe Val AAA TAC ACA CTG TTC CAA ATC TTC AGC AAG ATC GAC CGC CCT GAA GCC TCC CGC ATC GCC CTG CTC CTG ATG GCC AGC CAG GAG CCC CAA CGG ATG TCC CGG AAC TTT GT 74 221 111 332 Arg Tyr Val Gin Giy Leu Lys Lys Lys Lys Val Ile Val Ile Pro Val Giy Ile Giy Pro His Ala Asn Leu Lys Gin Ile Arg Leu Ile Giu Lys Gin Ala Pro Giu Asn CGC TAC GIC CAG GGC CTG AAG AAG AAG AAG GIC ATT GIG ATC CCG GIG GGC ATT GGG CCC CAT GCC AAC CTC AAG CAG CTC GTC GAG GAC CCC GAG GAC Lys Ala Phe Val Leu Ser Ser Val Asp Giu Leu Giu Gin Gin Arg Asp Giu lle Val Ser Tyr Leu Cys Asp Leu Ala Pro Giu Ala Pro Pro Pro Thr Leu Pro Pro Asp AAG GCC TTC GTG CTG AGC AGT GTG GAT GAG CTG GAG CAG CAA AGG GAC GAG ATC GTT AGC TAC CTC TGT GAC CTT GCC CCT GAA GCC CT CCT CCT ACT CTG CCC CGAC 148 Met Ala Sin Val Thr Val Si Pro Gly Leu Leu Gly Val Ser Thr Leu Gly Pro Lys Arg Asn Ser Met Val Leu Asp Val Ala Phe Val Leu Glu Gly Ser Asp Lys 11 ATG GCA CAA GTC ACT GTG GGC CCG GGG CTC TTG GGS GTT TCG ACC CTG GGS CCC AAG AGG AAC TCC ATG GTT CTG GAT GTG GCG TTC GTC CTG GAA GGA TCG ACA AAT 185 554 GIY GIU AIA ASP PHE ASN Arg Ser Lys GIU PHE MET GIU GIU VAI IIE GIN Arg MET ASP VAI GIY GIN ASP Ser IIE HIS VAI THY VAI LEU GIN TYY SER TYY MET VAI GGT GAA GCC GAC TIC AAC AGG AGC AAG GAG TIC ATG GAG GAG GTG ATT CAG CGG ATG GAG GAG GAG GAC AGC ATC CAC GTC ACG GTG CTG CAG TAC TCC TAC ATG GTG 222 665 Thr Val Glu Tyr Pro Phe Ser Glu Ala Gln Ser Lys Gly Asp lle Leu Gln Arg Val Arg Glu Ile Arg Tyr Gln Gly Gly Asn Arg Thr Asn Thr Gly Leu Ala Leu Arg ACC GTG GAG TAC CCC TTC AGC GAG GCA CAG TCC AAA GGG GAC ATC CTG CAG CGG GTG CGA GAG ATC CGC TAC CAG GGC GAC AGC ACC AGC ACT GGG CTG GCC CTG CGG 259 776 296 887 Tyr Leu Ser Asp His Ser Phe Leu Val Ser Gin Giy Asp Arg Giu Gin Ala Pro Asn Leu Val Tyr Met Val Thr Giy Asn Pro Ala Ser Asp Giu Ile Lys Arg Leu Pro TAC CTC TCT GAC CAC AGC TTC TTG GTC AGC CAG GGT GAC CGG GAG CAG GCG CCC AAC CTG GTC TAC GTC ACC GGA AAT CCT GCC TCT GAT GAG ATC AAG AGG CTG CCT 333 GIY ASP ILE GIN VAL VAL PRO ILE GIY VAL GIY PRO ASN ALA ASN VAL GIN GIU LEU GIU ARG ILE GIY TRP PRO ASN ALA PRO ILE LEU ILE GIN ASP PHE GIU THR LEU GGA GAC ATC CAG GIG GIG CCC ATT GGA GIG GGC CCT AAT GCC AAC GIG CAG GAG GIG GAG AGG AGT GGC IGG CCC AAT GCC CT ATC CAG GAC TIT GAG AGC TCC Pro Arg Glu Ala Pro Asp Leu Val Leu Gln Arg Cys Cys Ser Gly Glu Gly Leu Gln 11e Pro Thr Leu Ser Pro Ala Pro Asp Cys Ser Gln Pro Leu Asp Val 11e Leu CCC CGA GAG GCT CCT GAC CTG GTG CTG CAG AGG TGC TGC TGC GGA GAG GGG CTG CAG ATC CCC ACC CTC CCC AGCA CCT GAC TGC AGC CAG CCC CTG GAC GTG ATC CTT 370 1109 407 1220 Val Leu Gin Tyr Gly Ser ile Thr Thr ile Asp Val Pro Trp Asn Val Val Pro Glu Lys Ala His Leu Leu Ser Leu Val Asp Val Met Gin Arg Glu Gly Gly Pro Ser GTG CTG CAG TAT GGA AGC ATC ACC ACT GAC GTG GCCA TGG AAC GTG GTC CCG GAG AAA GCC CAT TTG CTG AGC CTT GTG GAC GTC ATG CAG CGG GAG GGA GGA GGC CCC AGC 444 1331 481 1442 GIN ILE GIY ASP ALA LEU GIY PHE ALA VAL AND TYN LEU THN SEN GIU MET HIS GIY ALA AND PNO GIY ALA SEN LYS ALA VAL VAL ILE LEU VAL THN ASP VAL SEN VAL CAA ATC GGG GAT GCC TTG GGC TIT GCT GTG CGA TAC TTG ACT TCA GAA ATG CAT GGT GCC AGG CGG GGA GCC TCA AAG GCG GTG GTC ATC CTG GTC ACG GAC GTC TTG GTG 518 1553 Asp Ser Val Asp Ala Ala Ala Asp Ala Ala Arg Ser Asn Arg Val Thr Val Phe Pro Ile Gly Ile Gly Asp Arg Tyr Asp Ala Ala Gin Leu Arg Ile Leu Ala Gly Pro GAT FCA GIG GAT GCA GCA GCT GAT GCC CAG FCC A AC GAG GTA GCA GCA GTT GTT GGA AT CGC TAC GTA GCA GCC AG GTA GGA GT 555 1664 Ala Gly Asp Ser Asn Val Val Lys Leu Gin Arg Ile Glu Asp Leu Pro Thr Met Val Thr Leu Gly Asn Ser Phe Leu <u>His Lys Leu Cys Ser Gly</u> Phe Val Arg Ile Cys GCA GGC GAC TCC AAC GTG GTG AAG CTC CAG CGA ATC GAA GAC CTC CCT ACC ATG GTC ACC TTG GGC AAT TCC TTC CTC CAC AAA CTG TGC TCT GGA TTT GTT AGG ATT TGC Met Asp Glu Asp Gly Asn Glu Lys Arg Pro Gly Asp Val Trp Thr Leu Pro Asp Gln Cys His Thr Val Thr Cys Gln Pro Asp Gly Gln Thr Leu Leu Lys Ser His Arg ATG GAT GAG GAT GGG AAT GAG AAG AGG CCC GGG GAC GTC TGG ACC TTG CCA GAC CAG TGC CAC GTG ACT TGC CAG GAC CTG CAG AGC TTG CTG AAG AGT CAT CGG Val Asn Cys Asp Arg Gly Leu Arg Pro Ser Cys Pro Asn Ser Gln Ser Pro Val Lys Val Glu Glu Thr Cys Gly Cys Arg Tro Thr Cys Pro Cys Val Cys Thr Gly Ser GTC AAC TGT GAC CGG GGG CTG AGG CCT TCG TGC CCT AAC AGC CAG TCC CCT GTT AAA GTG GAA GAG ACC TGT GGC TGC GCG TGG ACC TGC GTG GTG ACA GAG AGC Ser Thr Arg His Ile Val Thr Phe Asp Gly Gln Asn Phe Lys Leu Thr Gly Ser Cys Ser Tyr Val Leu Phe Gln Asn Lys Glu Gln Asp Leu Glu Val Ile Leu His Asn TCC ACT CGG CAC ATC GTG ACC TTT GAT GGG CAG AAT TTC AAG CTG ACT GGC AGC TGT TCT TAT GTC CTA TTT CAA AAC AAG GAG CAG GAC CTG GAG GTG ATT CTC CAT AAT 703 2108 Gly Ala Cys Ser Pro Gly Ala Arg Gln Gly Cys Met Lys Ser lle Glu Val Lys His Ser Ala Leu Ser Val Glu Leu His Ser Asp Met Glu Val Thr Val Asn Gly Arg GGT GCC TGC AGC CCT GGA GCA AGG CAG GGC TGC ATG AAA TCC ATC GAG GTG AAG CAC AGT GCC CTC TCC GTC GAG CTG CAC AGT GGA GTG ACG GTG ACT GGG AGA Leu Val Ser Val Pro Tyr Val Gly Gly Asn Met Glu Val Asn Val Tyr Gly Ala 11e Met His Glu Val Arg Phe Asn His Leu Gly His 11e Phe Thr Phe Thr Pro Gln CTG GTC TCT GTT CCT TAC GTG GGT GGG AAC ATG GAA GTC AAC GTT TAT GGT GCC ATC ATG CAT GAG GTC AGA TTC AAT CAC CTT GGT CAC ATC TCT CACA TTC ACA T Asn Asn Glu Phe Gln Leu Gln Leu Ser Pro Lys Thr Phe Ala Ser Lys Thr Tyr Gly Leu Cys Gly Ile Cys Asp Glu Asn Gly Ala Asn Asp Phe Met Leu Arg Asp Gly AAC AAT GAG TTC CAA CTG CAG CTC AGC CCC AAG ACT TTT GCT TCA AAG ACG TAT GGT CTG TGT GGG ATC TGT GAG GAC GGA GCC AAT GAC TTC ATG CTG AGG GAT GGC 777 2330 814 2441 Thr Vai Thr Thr Asp Trp Lys Thr Lev Vai Gin Giu Trp Thr Vai Gin Arg Pro Gly Gin Thr Cys Gin Pro I Lev Giu Giu Gin Cys Lev Vai Pro Asp Ser Ser His AcA GTC ACC ACA GAC GG6 AAA ACA CTI GTI CAG GAA TGC ATT A GTG ACT GTG CAG GAA GAC TGT CTC CAG AGA GTC TAGA Cys Gln Val Leu Leu Pro Leu Phe Ala Glu Cys His Lys Val Leu Ala Pro Ala Thr Phe Tyr Ala Ile Cys Gln Gln Asp Ser Cys His Gln Glu Gln Val Cys Glu TGC CAG GTC CTC CTC TTA CCA CTG TTT GCT GAA TGC CAC AAG GTC CTG GCT CCA GCC ACA TTC TAT GCC ATC TGC CAG GAC AGT TCG CAC CAG GAG GAA GTG TGT GAG 851 2552 Val lle Ala Ser Tyr Ala His Leu Cys Arg Thr Asn Gly Val Cys Val Asp Trp Arg Thr Pro Asp Phe Cys Ala Met Ser Cys Pro Pro Ser Leu Val Tyr Asn His Cys GTG ATC GCC TCT TAT GCC CAC CTC TGT CGG ACC AAC GGG GTC TGC GTT GAC TGG AGG ACA CCT GAT TTC TGT GCT ATG TCA TGC CCA CCA TCT CTG GTC TAC AAC CAC TGT 888 2663 925 2774 Glu His Gly Cys Pro Arg His Cys Asp Gly Asn Val Ser Ser Cys Gly Asp His Pro Ser Glu Gly Cys Phe Cys Pro Pro Asp Lys Val Met Leu Glu Gly Ser Cys Val GAG CAT GGC TGT CCC CGG CAC TGT GAT GGC AAC GTG AGC TCC TGT GGG GAC CAT CCC TCC GAA GGC TGT TTC TGC CCT CCA GAT AAA GTC ATG TTG GAA GGC AGC TGT GTC Pro Glu Glu Ala Cys Thr Gln Cys lle Gly Glu Asp Gly Val Gln His Gln Phe Leu Glu Ala Trp Val Pro Asp His Gln Pro Cys Gln lle Cys Thr Cys Leu Ser Gly CCT GAA GAG GCC TGC ACT CAG TGC ATT GGT GAG GAT GGA GTC CAG CAC CAG TTC CTG GAA GCC TGG GTC CCG GAC CAC CAG CCC TGT CAG ATC TGC ACA TGC CTC AGC GGG 962 2885 Arg Lys Val Asn Cys Thr Thr Gln Pro Cys Pro Thr Ala Lys Ala Pro Thr Cys Gly Leu Cys Glu Val Ala Arg Leu Arg Gln Asn Ala Asp Gln Cyc Cys Pro Glu Tyr CGG AAG GTC AAC TGC ACA ACG CAC CGC CCC ACG GCC AAA GCT CCC ACG TGT GGC CTG TGT GAA GTA GCC CGC CTC CGC CAG AAT GCA GAC CAG TGC TGC CCC GAG TAT Glu <u>Cys Val Cys Asp Pro Va</u>l Ser Cys Asp Leu Pro Pro Val Pro His Cys Glu Arg Gly Leu Gln Pro Thr Leu Thr Asn Pro Gly Glu Cys Arg Pro Asn Phe Thr Cys GAG TGT GTG TGT GAC CCA GTG AGC TGT GAC CTG CCC CCA GTG CCT CAC TGT GAA CGT GGC CTC CAG CCC ACA CTG ACC CAG CGG GGC GAG AGC CCA AC TTC ACC TGC 1036 3107 1073 3218 1110 3329 Cys Val Asn Ser Thr Val Ser Cys Pro Leu Gly Tyr Leu Ala Ser Thr Ala Thr Asn Asp Cys Gly Cys Thr Thr Thr Thr Cys Leu Pro Asp Lys Val Cys Val His Arg TGT GTC AAC TCC ACA GTG AGC TGT CCC CTT GGG TAC TTG GCC TCA ACC GUC ACC GAC AAT GAC TGT GGC TGT ACC ACA ACC TGC CTT CCC GAC AAG GTG TGT CTC CAC CGA 1147 3440 Ser Thr 11e Tyr Pro Val Gly Gln Phe Trp Glu Glu Gly Cys Asp Val Cys Thr Cys Thr Asp Met Glu Asp Ala Val Met Gly Leu Arg Val Ala Gln Cys Ser Gln Lys Age Acc Art TA C CT GT GG CAG THT TG G GAG GAG GG CT GG GAT GTG TG CAC CT GC GAC ATG GAG GAT GGC GGC GTG GTG GG GC CAG TGC TC C CAG AGA Pro Cys Glu Asp Ser Cys Arg Ser Gly Phe Thr Tyr Val Leu His Glu Gly Glu Cys Cys Gly Arg Cys Leu Pro Ser Ala Cys Glu Val Val Thr Gly Ser Pro Arg Gly CCC TGT GAG GAC AGC TGT CGG TCG GGC TTC ACT TAC GTT CTG CAT GAA GGC GAG TGC TGT GGA AGG TGC CTG CCA TCT GCC TGT GAG GTG GTG ACT GGC TCA CCG CGG 1184 3551 ASD SET GIN SET SET TED LYS SET VAI GIY SET GIN TED AIA SET PTO GIU ASN PTO CYS LEU IIE ASN GIU CYS VAI AFG VAI LYS GIU GIU VAI PHE IIE GIN GIN AFG GAC TCC CAG TCT TCC TGG AAG AGT GTC GGC TCC CAG TGG GCC TCC CCG GAG AAC CCC TGC CTC ATC AAT GAG TGT GTC CGA GTG AAG GAG GAG GTC TTT ATA CAA CAA AGG 1221 3662 1258 3773 1295 3884 Met Leu Asn Gly Thr Val Ile Gly Pro Gly Lys Thr Val Met Ile Asp Val Cys Thr Thr Cys Arg Cys Met Val Gln Val Gly Val Ile Ser Gly Phe Lys Leu Glu Cys ATG CTC AAT GGC ACT GTC ATT GGG CCC GGG AAG ACT GTG ATG ATG GTG GTG GTG AGG ACC TGC CGC TGC ATG GTG GGG GTG ATG CT GGA TTC AAG CTG GAG TGC Arg Lys Thr Thr Cys Asn Pro Cys Pro Leu Gly Tyr Lys Glu Glu Ašn Asn Thr Gly Glu Cys Cys Gly Arg Cys Leu Pro Thr Ala Cys Thr 11e Gln Leu Arg Gly Gly AGG AAG ACC ACC TGC AAC CCC TGC CCC CTG GGT TAC AAG GAA GAA AAT AAC ACA GGT GAA TGT TGT GGG AGA TGT TTG CCT ACG GCT TGC ACC ATT CAG CTA AGA GGA GGA 1332 3995 GIN ILE MET THY LEW LYS ANG ASD GIU THY LEW GIN ASD GIY CYS ASD THY HIS PHE CYS LYS VAI ASN GIU ANG GIY GIU TYY PHE TYP GIU LYS ANG VAI THY GIY CYS CAG ATC ATG ACA CTG AAG CGT GAT GAG ACG CTC CAG GAT GGC TGT GAT ACT CAC TTC TGC AAG GTC AAT GAG AGA GAG GAC TAC TTC TGG GAG AAG AGG GGT ACA GGC TGC 1369 4106 Pro Pro Phe Asp Glu His Lys Cys Leu Ala Glu Gly Gly Lys Ile Met Lys Ile Pro Gly Thr Cys Cys Asp Thr Cys Glu Glu Pro Glu Cys Asn Asp Ile Thr Ala Arg CCA CCC TTT GAT GAA CAC AAG TGT CTG GCT GAG GGA GGT AAA ATT ATG AAA ATT CCA GGC ACC TGC TGT GAC ACA TGT GAG GAG CCT GAG TGC AAC GAC ATC ACT SCC AGG 1406 4217 Lew Gin Tyr Val Lys Val Giy Ser Cys Lys Ser Giu Val Giu Val Asp lie His Tyr Cys Gin Giy Lys Cys Ala Ser Lys Ala Met Tyr Ser lie Asp lie Asn Asp Val CTG CAG TAT GTC AAG GTG GGA AGC TGT AAG TCT GAA GTA GAG GTG GAT ATC CAC TAC TGC CAG GGC AAA TGT GCC AGC AAA GCC ATG TAC TCC ATT GAC ATC AAC GAT GTG GIN ASP GIN CYS SER CYS CYS SER PRO THR AFG THR GIU PRO MEE GIN VAI AIA LEU HIS CYS THR ASN GIY SER VAI VAI TYR HIS GIU VAI LEU ASN AIA MEE GIU CYS CAG GAC CAG TGC TCC TGC TGC TCC CG ACA CGG ACG GAG GCC CATG CAG GTG GCC CTG CAC TGC ACT GGC TCT GTG TAC CAT GAG Lys Cys Ser Pro Arg Lys Cys Ser Lys STOP AAA TGC TCC CCC AGG AAG TGC AGC AAG TGA GGCTGCTGCA GCTGCATGGG TGCCTGCTGC TGCCTGCCTT GGCCTGATGG CCAGGCCAGA GTGCTGCCAG TCCTCTGCAT GTTCTGCTCT 4670 TGTGCCCTTC TGAGCCCACA ATAAAGGCTG AGCTCTTATC TTGC...(A)

FIG. 2. Nucleotide sequences for the cDNA inserts in λ HvWF1 and λ HvWF3 and predicted amino acids. Nucleotides are numbered 1-404 in λ HvWF1 and 1-4713 in λ HvWF3, excluding the poly(A) tail. The serine, which is the amino-terminal residue of von Willebrand factor present in plasma, has been numbered as residue 1 in λ HvWF1. The numbering of the amino acids in λ HvWF3 (1-1525) is arbitrary and will require revision when the DNA sequence separating these two isolates is established. Asparagine residues that are potentially glycosylated are shown by solid diamonds. Amino acids that are present in repeated segments are underlined. The beginning and ending of the A domains are identified by solid circles, the B domains are shown by solid triangles, and the C domains are shown by solid squares. The tetrapeptide, Arg-Gly-Asp-Ser, is identified by arrowheads; the RNA processing or polyadenylylation signal of AATAAA is enclosed in a box.

Biochemistry: Sadler et al.

DOMAIN Repeat Repeat Repeat	A 1 2 3	triplica 1 192 385	tio P E	n GLLC GL-(V S I P	ILGI ILSI	PKR PAP	N-S DCS	M VL Q PL	D V A D V I	F VL L LL	EGS DGS	D K 1 S S F	SE GE PA	E F E D F N S Y F C	VLI IRSI Demi	C A F V C E F M C S F A	V D I E E K A	M M É I VIQI FISI	R L R R M D V K A N	ISQK VGQD IGPR	W V R V A S I H V T L T Q V S	29 250 443	
Repeat Repeat Repeat	1 2 3	30 251 444	V V V	V EY L Q Y L Q Y	1 DG] G	S S Y M S	H VTV ITT	AYI EY- I D V	GLK PFS PWW	D R K E A Q I V V P	R P S S K G E K A	E-L DIL HLL	RRI QRV SIV	(AS) /-R /DV	QVKY EIRY MQR-	AG QG EG	S Q V - G W R - G P S Q) I G	A STE T NT (D A L	S E VI Glai Gfai	L R Y - L R Y - V R Y -	L L S D H S L T	77 301 494	
Repeat Repeat Repeat	1 2 3	78 302 495	F	Q I F L V -	S K 1 S Q G S E M	D D H G A	R PE RE Q R PG	A SR AP M A SK	- I A - L V A V V	LLL V MV ILV	MAS TG- TDV	Q E P S V D	Q Ř I – N I 9 S V I	ISR PAS DAA	NFVE DEIE ADAZ	RL R-	26L1 PG	K K - S N		VIP VVP VFP	VG I G I GVG I G I G	PHANL PNANV DRYDA	134 348 545	
Repeat Repeat Repeat	1 2 3	135 349 546	K Q A	QIR Ele Ql-	LIE RI- RIL	K G W P A G P	Q A P N A P A G D	ENK SN-	A F 1 [/ L S S L / V - K	V D E IQE LQE	LEQ FE- IE-	Q R I 	D E İ 	V S Y I 	, cD - T - D	LA PE L PRE L PT P	E A P E A P I V T	P P T D L G N	L P P I L V S F	D М А Q - L Q R - L H K	V T VG CC S G LC S G	191 384 586	
DO MAIN Repeat Repeat	B 1 2	duplica 1008 1087		n PTA PPH	K A P R L P	TCG T	L C E 	V A B	LR) N A D (T	Q C C	PEY DEY	EC	VCD NCN	PV 10 CV 11)42 11								
DOMAIN Repeat Repeat	C 1 2	duplica 1112 1256	tio N	n STV RNV	S C P S C P	- L - Q L E	 V P V	 C P S	GY 1 GF (LAST LSC	AT N Kts	IDCG AC-	CT CP	T T T S C R		D K V	сигі сигі	R ST N GT	I Y P V I G	VGQ PGK	- F H E T V M 1	EGCDV DVCTT	CT CR	1164 1316
Repeat Repeat	1	1165 1317		T D M M V Q	eda VgV	V M G I S G	L R V F K L	А Q (С - в (С	S Q I R K 1	CPCP TC-	DSC NPC	R S C P L C	G F T G Y K	Y V L E E N	H E G I N T G I	E C C E C C	G R C I G R C I	L P S L P T	A C B A C T	VVŤ IQL	G S P S S	GDSQS GGQIM	S WI T L	(1226 (1374

FIG. 3. Internal homologous domains in von Willebrand factor. The amino acids are numbered as shown in Fig. 2 for λ HvWF3. The alignments shown were achieved with the ALIGN computer program (40). For each pairwise comparison, the alignment scores in standard deviation units (SD) are followed by the probability (P) that this score could occur by chance: Domain A repeat 1 versus 2, 5.96 SD, $P < 10^{-8}$; repeat 1 versus 3, 3.29 SD, $P < 10^{-3}$; repeat 2 versus 3, 10.62 SD, $P < 10^{-23}$. Domain B repeat 1 versus 2, 5.98 SD, $P = 10^{-9}$. Domain C repeat 1 versus 2, 7.76 SD, $P < 10^{-13}$. The scoring matrix used was the mutation data matrix (MD + 2), with a gap penalty of 6 (40). Amino acids are shown by the single-letter code.

be due to polymorphism in the protein or a cloning artifact generated during the preparation of the cDNA library.

von Willebrand factor contains $\approx 15\%$ carbohydrate distributed among both asparagine-linked and threonine/serinelinked oligosaccharides. Thus far, at least five asparaginelinked structures have been identified (42, 43). The protein sequence predicted from λ HvWF1 and λ HvWF3 includes 11 potential *N*-glycosylation sites with the sequence of Asn-X-Thr(or Ser). In addition, there are seven potential carbohydrate binding sites with the sequence of Asn-X-Cys. This sequence is glycosylated in protein C (44, 45). The carboxylterminal one-third of von Willebrand factor is very rich in cysteine residues, while the methionine residues are distributed throughout the molecule.

Three separate polypeptide segments of von Willebrand factor show evidence of internal duplication (Fig. 3). Amino acid residues 1-586 in λ HvWF3 contain a head-to-tail triplication of ≈ 200 amino acids with an identity ranging from 29%to 43%. The amino acid sequence between residues 1008 and 1042 is duplicated in residues 1087-1111. Also, the amino acid sequence between residues 1112 and 1226 is duplicated in residues 1256-1374. For each comparison, the alignment score in standard deviation units is listed in the figure legend. Among the three groups of duplicated amino acid sequences, there were small regions that show remarkable sequence identity and other regions that have diverged considerably. Altogether, $\approx 50\%$ of the amino acid sequence shown in Fig. 2 is part of repeated segments. The function of these regions is unknown, but they do indicate that von Willebrand factor has a complex evolutionary history involving gene duplication of the three different sequences or domains.

Nucleotides 3656-3667 in λ HvWF3 code for Arg-Gly-Asp-Ser. This sequence has been identified within a region of fibronectin that mediates the cell attachment activity of the protein (46, 47). In addition, both von Willebrand factor and fibronectin bind competitively to thrombin-activated platelets, and these interactions are inhibited by small peptides containing the sequence Arg-Gly-Asp-Ser. This suggests that this region of von Willebrand factor may participate in platelet binding (48, 49). von Willebrand factor, however, is not homologous to fibronectin. In fact, a comparison of the von Willebrand factor sequence with those in the National Biomedical Research Foundation Protein Sequence Data Base (Georgetown University, Washington, DC; ref. 40) did not reveal any major similarity to fibronectin or any other protein. If this region of von Willebrand factor is required for a physiologically significant interaction with platelets, the similarity to the corresponding functional region of fibronectin may represent convergent evolution.

We thank Drs. Kenneth Walsh, Kazuo Fujikawa, Santosh Kumar, and Lowell Ericsson for helpful discussions dealing with the protein sequence of human von Willebrand factor; Mark Behlke for performing Maxam-Gilbert sequencing reactions on selected fragments; John Bell for synthesis of oligonucleotides; Penny Thompson for culturing the endothelial cells; and Dr. Michael Chopek for advice and for kindly providing purified von Willebrand factor. Also, the assistance of Drs. Mark Murray and Fred Hagen in constructing the cDNA library in λ gt11 is gratefully acknowledged. This work was supported in part by National Institutes of Health Research Grants HL 16919, HL 18645, and HL 29595. J.E.S. was an Associate of the Howard Hughes Medical Institute, University of Washington, Seattle (1982–1984), and is currently an Associate Investigator of the Howard Hughes Medical Institute, Washington University, St. Louis.

- Jaffe, E. A., Hoyer, L. W. & Nachman, R. L. (1973) J. Clin. Invest. 52, 2757-2764.
- Jaffe, E. A. & Hoyer, L. W. (1974) Proc. Natl. Acad. Sci. USA 71, 1906–1909.
- Nachman, R. L., Levine, R. & Jaffe, E. A. (1977) J. Clin. Invest. 60, 914-921.

- 4. Okumura, T. & Jamieson, G. A. (1976) Thromb. Res. 8, 701-706.
- Nachman, R. L., Jaffe, E. A. & Weksler, B. B. (1977) J. Clin. Invest. 59, 143-148.
- Jenkins, C. S. P., Phillips, D. R., Clemetson, K. J., Meyer, D., Larrieu, M.-J. & Luscher, E. F. (1976) J. Clin. Invest. 57, 112-124.
- Fujimoto, T., Ohara, S. & Hawiger, J. (1982) J. Clin. Invest. 69, 1212–1222.
- Ruggeri, Z. M., Nilsson, I. M., Lombardi, R., Holmberg, L. & Zimmerman, T. S. (1982) J. Clin. Invest. 70, 1124-1127.
 Santoro, S. A. (1981) Thromb. Res. 21, 689-693.
- Santoro, S. A. & Cowan, J. F. (1982) Collagen Relat. Res. 2, 31–34.
- 11. Morton, L. F., Griffin, B., Pepper, D. S. & Barnes, M. J. (1983) Thromb. Res. 32, 545-556.
- Tuddenham, E. G. D., Lane, R. S., Rotblat, F., Johnson, A. J., Snape, T. J., Middleton, S. & Kernoff, P. B. A. (1982) Br. J. Haematol. 52, 259-267.
- 13. Wagner, D. D. & Marder, V. J. (1983) J. Biol. Chem. 258, 2065-2067.
- Lynch, D. C., Williams, R., Zimmerman, T. S., Kirby, E. P. & Livingston, D. M. (1983) Proc. Natl. Acad. Sci. USA 80, 2738-2742.
- 15. Ling, E. H., Browning, P. J., Zimmerman, T. S. & Lynch, D. C. (1984) Circulation 70, Suppl. 2, abstr. 833.
- Wagner, D. D. & Marder, V. J. (1984) J. Cell Biol. 99, 2123-2130.
- 17. Browning, P. J., Ling, E. H., Zimmerman, T. S. & Lynch, D. C. (1983) Blood 62, 281 (abstr.).
- Sixma, J. J., Sakariassen, K. S., Stel, H. V., Houdijk, W. P. M., In der Maur, D. W., Hamer, R. J., de Groot, P. G. & van Mourik, J. A. (1984) *J. Clin. Invest.* 74, 736-744.
- Titani, K., Ericsson, L. H., Kumar, S., Dörsam, H., Chopek, M. W. & Fujikawa, K. (1984) Circulation 70, Suppl. 2, abstr. 837.
- Girma, J. P., Pietu, G., Chopek, M. W., Edgington, T. S. & Meyer, D. (1984) *Circulation* 70, Suppl. 2, abstr. 836.
- Hoyer, L. W. (1982) in *The Hemophilias*, ed. Bloom, A. L. (Churchill-Livingstone, Edinburgh, Scotland), pp. 106–121.
- 22. Kinoshita, S., Harrison, J., Lazerson, J. & Abildgaard, C. F. (1984) *Blood* 63, 1369-1371.
- Sadler, J. E., Titani, K., Harlan, J. M. & Davie, E. W. (1985) Fed. Proc. Fed. Am. Soc. Exp. Biol. 44, 1069 (abstr.).
- Lynch, D. C., Zimmerman, T. S., Collins, C. J., Morin, M. J., Ling, E. H. & Livingston, D. M. (1985) *Clin. Res.* 33, 548 (abstr.).
- Ginsburg, D., Handin, R. I., Bonthron, D. T. & Orkin, S. H. (1985) Clin. Res. 33, 546 (abstr.).
- Striker, G. E., Harlan, J. M. & Schwartz, S. M. (1980) Methods Cell Biol. 21, 135-151.
- Thornton, S. C., Mueller, S. N. & Levine, E. M. (1983) Science 221, 623-625.

- Chirgwin, J. M., Przbyla, A. E., MacDonald, R. J. & Rutter, W. J. (1979) *Biochemistry* 18, 5294-5299.
- Schwarzbauer, J. E., Tamkun, J. W., Lemischka, I. R. & Hynes, R. O. (1983) Cell 35, 421-431.
- 30. Canfield, W. M. & Kisiel, W. (1982) J. Clin. Invest. 70, 1260-1272.
- Salacinski, P. R. P., McLean, C., Sykes, J. E. C., Clement-Jones, V. V. & Lowry, P. J. (1981) Anal. Biochem. 117, 136-146.
- 32. Young, R. A. & Davis, R. W. (1983) Science 222, 778-782.
- Maniatis, T., Fritsch, E. F. & Sambrook, J. (1982) Molecular Cloning: A Laboratory Manual (Cold Spring Harbor Laboratory, Cold Spring Harbor, NY), pp. 75-85.
- Degen, S. J., MacGillivray, R. T. A. & Davie, E. W. (1983) Biochemistry 22, 2087-2097.
- 35. Norrander, J., Kempe, T. & Messing, J. (1983) Gene 26, 101-106.
- Sanger, F., Nicklen, S. & Coulson, A. R. (1977) Proc. Natl. Acad. Sci. USA 74, 5463-5467.
- Biggin, M. D., Gibson, T. J. & Hong, G. F. (1983) Proc. Natl. Acad. Sci. USA 80, 3963–3965.
- Poncz, M., Solowiejczyk, D., Ballantine, M., Schwartz, E. & Surrey, S. (1982) Proc. Natl. Acad. Sci. USA 79, 4298-4302.
- 39. Maxam, A. M. & Gilbert, W. (1980) Methods Enzymol. 65, 499-560.
- Dayhoff, M. O., Barker, W. C. & Hunt, L. T. (1983) Methods Enzymol. 91, 524-545.
- Blobel, G., Walter, P., Chang, C. N., Goldman, B. M., Erickson, A. H. & Lingappa, R. (1979) in Secretory Mechanisms, Society for Experimental Biology Symposium, eds. Hopkins, C. R. & Duncan, D. J. (Cambridge University Press, London), Vol. 33, pp. 9–36.
- Samor, B., Mazurier, C., Goudemand, M., Debeire, P., Fournet, B. & Montreuil, J. (1982) Thromb. Res. 25, 81-89.
- Debeire, P., Montreuil, J., Samor, B., Mazurier, C., Goudemand, M., Van Halbeek, H. & Vliegenthart, J. F. G. (1983) FEBS Lett. 151, 22-26.
- 44. Stenflo, J. & Fernlund, P. (1982) J. Biol. Chem. 257, 12180-12190.
- 45. Foster, D. & Davie, E. W. (1984) Proc. Natl. Acad. Sci. USA 81, 4766-4770.
- 46. Pierschbacher, M. D. & Ruoslahti, E. (1984) Nature (London) 309, 30-33.
- Pierschbacher, M. D. & Ruoslahti, E. (1984) Proc. Natl. Acad. Sci. USA 81, 5985-5988.
- Ginsburg, M., Pierschbacher, M. D., Ruoslahti, E., Marguerie, G. & Plow, E. (1985) J. Biol. Chem. 260, 3931–3936.
- Haverstick, D. M., Cowan, J. F., Yamada, K. M. & Santoro, S. A. (1985) Blood, in press.
- Hessel, B., Jörnvall, H., Thorell, L., Söderman, S., Larsson, U., Egberg, N., Blombäck, B. & Holmgren, A. (1984) Thromb. Res. 35, 637-651.