cDNA Sequences for Human von Willebrand Factor Reveal Five Types of Repeated Domains and Five Possible Protein Sequence Polymorphisms[†]

Beverley B. Shelton-Inloes, * Koiti Titani, * and J. Evan Sadler*, * *

Howard Hughes Medical Institute Laboratories, Departments of Medicine and Biochemistry, Washington University School of Medicine, St. Louis, Missouri 63110, and Department of Biochemistry, University of Washington, Seattle, Washington 98195

Received January 14, 1986; Revised Manuscript Received April 11, 1986

ABSTRACT: A human umbilical vein endothelial cell cDNA library in λgt11 was screened with two previously described cDNA inserts for human von Willebrand factor. Among 16 positive isolates, two that hybridized with a probe corresponding to the amino terminus of von Willebrand factor were sequenced. Together, these four cDNA inserts span 6.5 kilobases of the von Willebrand factor mRNA sequence, completely specifying the 2050 amino acids of the subunit of mature, secreted von Willebrand factor and 24 residues of a precursor peptide. Approximately 77% of the sequence is contained in five types of repeated domains. Domain A consists of 193-220 amino acids and is present in three tandem copies between residues 497 and 1111. Domain B contains 25-35 amino acids and is present in three copies between residues 1533 and 1636. Domain C consists of 116-119 amino acids and is duplicated between residues 1637 and 1899. In contrast to the essentially contiguous repetition of domains A-C, the two copies of domains D and E are each separated by 804 and 1383 amino acids, respectively. Domain D1 contains 289 amino acids between residues 79 and 367, while domain D2 consists of 270 amino acids between residues 1171 and 1440. Domain E1 consists of 46 amino acids between residues 25 and 70, and domain E2 consists of 46 amino acids between residues 1453 and 1498. The triplicated A domains are notably poor in Cys content, while the remaining domains are Cys-rich. The A domains appear to be homologous to a 225-residue segment of complement factor B. Otherwise, von Willebrand factor is not closely related to any protein in the National Biomedical Research Foundation Protein Sequence Database, nor is the portion of the von Willebrand factor cDNA sequence that encodes the secreted protein homologous to any sequence in the Genbank Genomic Sequence Data Bank. Thus, four of the five types of repeated domains in von Willebrand factor have no homologues among other known proteins. The tetrapeptide Arg-Gly-Asp-Ser occurs at the carboxy-terminal end of domain C1 and may mediate the binding of von Willebrand factor to the GPIIb/IIIa complex of activated platelets. All of domain A1 lies within a 50-kilodalton tryptic fragment of von Willebrand factor that binds to GPIb of resting platelets [Fujimura, Y., Titani, K., Holland, L. Z., Russell, S. R., Roberts, J. R., Elder, J. H., Ruggeri, Z. M., & Zimmerman, T. S. (1986) J. Biol. Chem. 261, 381-385]. The remaining domains (B, D, and E) have not been correlated with specific functions. The sequence of the von Willebrand factor precursor before the amino-terminal Ser of plasma von Willebrand factor is His-Arg-Ser-Lys-Arg-Ser. The mature subunit is generated by proteolytic cleavage after the Lys-Arg dipeptide. This sequence resembles that of several mammalian, viral, fungal, and yeast protein precursors that are also proteolytically processed after paired basic residues during biosynthesis. Among the four cDNA isolates sequenced by this laboratory, there are eight single-nucleotide discrepancies that may reflect polymorphism in the von Willebrand factor gene sequence. Of these, seven are transitions, and one is a transversion. Five do not affect the translated protein sequence, but four result in single amino acid substitutions. Together with the previously reported disagreement at residue 7 between the translated sequence of \(\lambda\)HvWF1 and the protein sequence, there are now five potential protein sequence polymorphisms for von Willebrand factor.

von Willebrand factor is a glycoprotein that is required for the adhesion of platelets to damaged endothelium. The concentration of von Willebrand factor in plasma is approximately $10 \,\mu g/mL$. It is also found in subendothelial connective tissue and platelet α granules. The form of the protein found in plasma contains a single type of subunit, with M_r 225 000, that forms oligomers ranging in size from dimers with M_r 500 000 to species with M_r over 12 000 000, held together by disulfide bonds [reviewed in Hoyer (1981)]. von Willebrand factor is synthesized by endothelial cells (Jaffe et al., 1973, 1974) and

by megakaryocytes (Nachman et al., 1977). The biosynthesis of von Willebrand factor is very complex and requires several posttransitional modifications, including proteolytic processing (Wagner & Marder, 1983; Lynch et al., 1983), glycosylation (Wagner & Marder, 1983, 1984), and sulfation (Browning et al., 1983).

von Willebrand factor is not an enzyme but participates in hemostatis through a variety of binding interactions. It forms a bridge between platelets and subendothelial connective tissue at sites of endothelial damage, and there are specific receptors for von Willebrand factor on each of these surfaces. In the presence of the antibiotic ristocetin, von Willebrand factor binds to glycoprotein Ib of resting platelets, causing platelet aggregation in plasma (Jenkins et al., 1976). The importance of this interaction is suggested by the bleeding diathesis that affects patients with the Bernard-Soulier syndrome, a disorder

[†]This work was supported in part by Research Grant HL29595 (to K.T.) from the National Institutes of Health. K.T. is a visiting professor of the Fujita-Gakuen Health University, Nagoya, Japan.

^{*}Correspondence should be addressed to this author.

¹Washington University School of Medicine.

[§] University of Washington.

characterized by the lack of functional glycoprotein Ib. von Willebrand factor binds to a second site on activated platelets, and binding to this site is competitive with the binding of fibronectin or fibrinogen (Pietu et al., 1984; Haverstick et al., 1985; Plow et al., 1985). The corresponding platelet receptor appears to be the glycoprotein IIb/IIIa complex. The target in subendothelial connective tissue may be collagen, since von Willebrand factor will bind specifically to purified fibrillar collagens but not to gelatin (Santoro, 1981; Santoro & Cowan, 1982; Morton et al., 1983). However, von Willebrand factor also binds to extracellular matrix that is devoid of detectable collagen, so that additional receptors may exist in the subendothelium (Wagner et al., 1984). Finally, factor VIII binds to von Willebrand factor and is stabilized in the circulation by this interaction. In the absence of von Willebrand factor, the survival of factor VIII is dramatically reduced, causing a secondary deficiency of factor VIII (Tuddenham et al., 1982). Consequently, patients with von Willebrand's disease, who lack functional von Willebrand factor, may exhibit bleeding characteristic either of platelet dysfunction or of classical hemophilia A.

Structure-function relationships of von Willebrand factor have been partly elucidated by protein chemistry, immunology, and molecular biology methods. Staphylococcal V8 protease separates native von Willebrand factor into two major fragments that are not connected by disulfide bonds. The amino-terminal fragment is a homodimer of 170-kilodalton peptides that retains the ability to bind to platelet glycoprotein Ib (Girma et al., 1986) and to collagen (Fressinaud et al., 1985). A 50-kilodalton tryptic fragment that also binds to platelet glycoprotein Ib has been placed within the linear sequence of this fragment (Fujimura et al., 1986). The carboxy-terminal V8 protease fragment is a homodimer of 100-kilodalton peptides that binds to activated platelets at the glycoprotein IIb/IIIa complex (Girma et al., 1984). Monoclonal antibodies to von Willebrand factor that selectively inhibit these binding activities have been shown to recognize still smaller fragments of von Willebrand factor produced by digestion with other proteases (Sixma et al., 1984). We have previously reported that a tetrapeptide segment, Arg-Gly-Asp-ser, occurs in the carboxy-terminal V8 protease fragment. The same sequence is required for the cell attachment and platelet binding activities of fibronectin (Pierschbacher & Ruoslahti, 1984a,b), and small peptides containing this sequence inhibit the binding of both fibronectin and von Willebrand factor to activated platelets (Haverstick et al., 1985; Plow et al., 1985). Thus, the domain that binds to the platelet glycoprotein IIb/IIIa complex probably contains this tetrapeptide.

In this paper, we describe the isolation of cDNA¹ inserts for von Willebrand factor that complete the nucleotide sequence corresponding to the species found in plasma. In addition, analysis of the translated amino acid sequence reveals the presence of two previously unrecognized duplicated domains within the sequence, and one additional copy of domain B (Sadler et al., 1985). There are a total of five unrelated repeated domains in the sequence, two of which are triplicated and three of which are duplicated. Together, these repetitive sequences comprise over three-fourths of the entire subunit.

The A domains may be homologous to a segment of complement factor B. The remaining domains do not have any homologue within other known protein sequences. In addition, the sequence of the von Willebrand factor precursor surrounding the amino terminus of the mature subunit is similar to that of several mammalian, viral, and yeast protein precursors, suggesting that the posttranslational proteolytic processing of these diverse proteins may proceed by similar mechanisms. Eight potential point mutations have been identified by comparing the nucleotide sequences of several cDNA isolates, of which four alter the predicted amino acid sequence.

MATERIALS AND METHODS

The human umbilical vein endothelial cell cDNA library in λ gt11, cDNA isolates λ HvWF1 and λ HvWF3, and methods for DNA preparation, subcloning, and sequencing have been described (Sadler et al., 1985). The reported nucleotide sequences were determined at least once on both strands. Deoxyadenosine 5'-[α -35S]thiotriphosphate ([α -32P]CTP) were purchased from Amersham. Deoxycytidine [α -32P]triphosphate ([α -32P]triphosphate ([α -32P]dCTP) was purchased from New England Nuclear.

cDNA Library Screening. The 404 base pair cDNA insert of λ HvWF1, corresponding to amino acids -24 to 110 of von Willebrand factor, was ligated into the EcoRI site of pGEM-1 (Riboprobe Gemini System, Promega-Biotec, Madison, WI), and the recombinant plasmid was propagated in E. coli HB101. Plasmid DNA was linearized with PvuII (or BamHI) and employed as a template for the SP6 (or T7) RNA polymerase to prepare RNA labeled with $[\alpha^{-32}P]$ CTP to a specific activity of 2×10^8 cpm/ μ g according to the instructions provided by the supplier.

A 2752 base pair FspI-SacI fragment, corresponding to amino acids 544-1461 of von Willebrand factor, was prepared by digestion of a SacI-SqcI subclone of $\lambda HvWF3$ in pUC18 (Sadler et al., 1985). The fragment was purified by polyacrylamide gel electrophoresis and electroelution and then labeled with $[\alpha^{-32}P]dCTP$ by nick translation (Maniatis et al., 1975) to a specific activity of $(1-2) \times 10^8$ cpm/ μg .

Approximately 1 500 000 recombinant phage from the λ gt11 cDNA library were plated on *E. coli* Y1088 at a density of 50 000 per 150-mm plate of LB agar and screened in duplicate by hybridization according to the method of Benton and Davis (1977) with both probes described above. Positive isolates were plaque-purified, and DNA was prepared for subcloning and sequencing.

Northern Blotting. Poly(A)+ RNA was prepared from cultured human umbilical vein endothelial cells (Sadler et al., 1985), and 5 μ g was employed for Northern blotting according to Thomas (1983). Size standards consisting of bovine liver 28S and 18S RNA and HindIII fragments of phage λ DNA were electrophoresed in adjacent lanes and stained with ethidium bromide. The blot was prehybridized in 7.5 M sodium citrate, 0.75 M sodium chloride, pH 7.0 (5× SSC), 50% (v/v) formamide, 5 mM sodium phosphate, 0.1% (w/v) Na-DodSO₄, 1 mM EDTA, 2.5× Denhart's solution, and 200 μg/mL denatured salmon sperm DNA, at 55 °C. The blot was hybridized in the same solution containing 500 000 cpm/mL of the \(\lambda\)HvWF1 probe (antisense) at 55 °C for 18 h, washed 3 times for 20 min at 65 °C in 0.1× SSC and 0.1% (w/v) NaDodSO₄, and exposed to Kodak XAR-5 film for 20 h.

Computer Analysis of Sequences. The von Willebrand factor protein sequence was compared to all entries in the NBRF Protein Sequence Database (Georgetown University,

¹ Abbreviations: NaDodSO₄, sodium dodecyl sulfate; EDTA, ethylenediaminetetraacetic acid; SSC, standard saline citrate (15 mM citrate, 0.15 M NaCl, pH 7.0); cDNA, complementary deoxyribonucleic acid; Denhardt's solution, 0.02% (w/v) bovine serum albumin, 0.02% (w/v) poly(vinylpyrrolidone), and 0.02% (w/v) Ficoll; NBRF, National Biomedical Research Foundation.

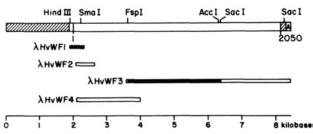


FIGURE 1: Restriction map of four human von Willebrand factor cDNA isolates. Only restriction sites used in subcloning and DNA sequencing are shown. In the summary map at the top of the figure, the hatched region to the left represents the leader peptide and 5'-noncoding sequences of the mRNA that are not represented in the cDNA isolates, and the hatched region at the right represents the 3'-noncoding sequence. The length of the complete von Willebrand factor cDNA (8.5 kilobases) is estimated from the Northern blot of Figure 2. The black regions of λHvWF1 and λHvWF3 indicate the fragments employed as hybridization probes in these studies. The scale at the bottom is in kilobases.

Washington, DC, release 6.0, August 28, 1985) and separately to the amino acid sequence of human fibronectin (Kornblihtt et al., 1985), with the programs Search, Relate, and Align (Dayhoff et al., 1983). The nucleotide sequence of the von Willebrand factor cDNA corresponding to amino acid residues -24 through 2050 was compared to all entries in the Genbank genomic sequence data bank (BBN Laboratories Inc., Cambridge, MA, release 38.0, November 11, 1985) with the program Fastn (Wilbur & Lipman, 1983). The amino acid sequence of plasma von Willebrand factor, residues 1-2050, was analyzed for internal segment duplications with the programs Relate and Align (Dayhoff et al., 1983).

RESULTS AND DISCUSSION

cDNA Library Screening and Sequence Comparisons. Among 1 500 000 recombinants screened, 16 hybridized with the λHvWF3 probe only, 1 with the λHvWF1 probe only (λHvWF2), and 1 with both probes (λHvWF4). The cDNA inserts from these two latter isolates were subcloned into M13mp18 for sequencing. A summary restriction map for the von Willebrand factor cDNA indicating the portion contained in these isolates is shown in Figure 1.

The cDNA insert of λHvWF2 consisted of 556 base pairs, encoding amino acids 22-206 of von Willebrand factor. A portion of the sequence overlapped with that determined previously for \(\text{AHvWF1 (Sadler et al., 1985).} \) The insert of λHvWF4 was 3.2 kilobases in length, and 1.9 kilobases at the 3' end corresponded to amino acids 18-661 of von Willebrand factor. Therefore, this insert overlapped with both λHvWF1 and hHvWF3, specifying the remaining amino acid sequence between these two clones (Figure 1). However, the sequence of the first 1.3 kilobases of this insert did not match that of λHvWF1, nor did the translated amino acid sequence match that determined by amino acid sequencing of von Willebrand factor. In addition, this segment contains no continuous open-reading frame, and there is no suitable acceptor splice junction sequence prior to the codon for amino acid 18 of von Willebrand factor. Thus, this segment may represent an unrelated DNA sequence that was juxtaposed inadvertently during construction or propogation of the cDNA library. Together, the four cDNA isolates span 6.5 kilobases of von Willebrand factor cDNA, including the poly(A) tail.

The length of the von Willebrand factor mRNA was estimated to be 8.5 kilobases by Northern blotting, as shown in Figure 2. This value is consistent with those obtained by others using similar methods (Lynch et al., 1985; Ginsburg et al., 1985; Verweij et al., 1985). Since only 6.4 kilobases

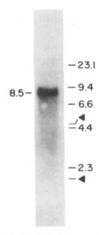


FIGURE 2: Northern blot of RNA from cultured human umbilical vein endothelial cells probed with a human von Willebrand factor cDNA probe. The mobility of 28S and 18S ribosomal RNA (A) and denatured DNA size markers (kilobases) is indicated to the right of the figure. The interpolated size of the autoradiographic signal is indicated on the left.

are necessary to encode the mature protein plus all 3'-non-coding mRNA sequences, 2 kilobases of mRNA might be available to encode a precursor or leader peptide as large as 700 amino acids, or 77 000 daltons (unglycosylated). On the basis of a comparison of amino-terminal protein sequence with translated von Willebrand factor cDNA sequence, Fay et al. (1985) have suggested that this precursor peptide is identical with von Willebrand antigen II, a 92 000–98 000-dalton (reduced) plasma protein of unknown function that is immunologically distinct from von Willebrand factor, originally described by Montgomery and Zimmerman (1978).

The sequences of isolates \(\lambda HvWF2 \) and \(\lambda HvWF4 \) are compared with those of \(\lambda\text{HvWF1}\) and \(\lambda\text{HvWF3}\) in Figure 3. This completes the nucleotide sequence of the cDNA corresponding to plasma von Willebrand factor. Compared to other human cDNA sequences, there are no striking bases in codon usage (Lathe, 1985). In the regions that overlap, there are eight single-nucleotide discrepancies, of which seven are transitions and one (\(\lambda\)HvWF4, nucleotide 1526, G to T) is a transversion. Four of the transitions affect the translated amino acid sequence. For these discrepancies, the predicted amino acid residue that agrees with the protein sequence (Titani et al., 1986) is underlined in Figure 3. We previously reported a discrepancy at amino acid residue 7 between λHvWF1 (His) and the protein sequence (Pro) (Sadler et al., 1985). The cDNA library employed for these studies was derived from a pool of 30-60 umbilical veins, and none of the tissue donors was known to harbor an abnormal von Willebrand factor allele. Thus, there are a total of five potential protein sequence polymorphisms implied by these cDNA sequences. One of these, Ala/Thr at residue 26, has been confirmed directly by protein sequencing. In von Willebrand factor prepared from pooled factor VIII concentrate, both Ala and Thr are identified at this position in a ratio of approximately 4:1 (Titani et al., 1986). At present, there is no way to exclude the possibility that the remaining discrepancies arose through errors of transcription during cDNA library preparation or that they represent nonfunctional mRNA sequences. A comparison of the sequence of $\lambda HvWF3$ with three other independently reported cDNA sequences does not show any additional discrepancies for the region encoding amino acids 1857 through the carboxy-terminus (Lynch et al., 1985; Ginsburg et al., 1985; Verweij et al., 1985). There are some disagreements between the sequence of the 3'-noncoding region

1080 GG TAT Gly Tyr	1170 TGT GTG Cys Val	1260 TGT GAG Cys Glu	1350 GTT GTC Val Val	1440 GTG GAG Val Glu	1530		1620	CAC G1u	1710	GCG Ala	1800	ATC 11e	18%)	ATF 1:e			
106 AAC GGG Asn Gly	CAG TGT Gln Cys	120 GTG TGT Val Cys	13: GAT GTT ASP VAL	14. TAT GTG Tyr Val		CTG TCC Leu Ser		Val	11	AAG TAT Lys Tyr	18	TCC CGC Ser Arg	=	AAG GTC Lys Val			
1070 CCC GAG Arg Glu	GTC Val	Pro Pro	1340 CAC TGT His Cys	£ 3		TCC AGG Ser Arg	1610	CTG CCC Val Ala	1700	CAG GTG Gln Val	06/1	GCC Ala		AAG AAG Lys Lys			
E 3	1160 C TGC CCT a Cys Pro	A GAC TGT u Asp Cys	2 %	460 1420 1430 6 GTG AGC CCC ACC ACT C o Val Ser Pro Thr Thr L 490	15	Ser	16	A'R	11	GCC AGC O		C CUT GAA g Pro Glu	81 .	CTC AAG A Leu Lys 1			
1060 GAG AGG AAT Glu Arg Asn 370	1150 CCA CTG GCC Pro Leu Ala 400	1240 GAC CCT GAA ASP Pro Glu 430	1330 TGC CAG ATT 1 Cys Gln Ile C	460 420 AGC CC Ser Pr 490	1510	GAT GGC Asp Gly 520	1600	TGC GTC Trp Val 550	0691	CGC ATT GC Arg 11e A1	1780	ATC CAC CGC CITE AS AND	1870	36.05	ŝ		
ose cac clu clu	CAG CCA Glu Pro	CTT CAC Val Asp	CAC TGC His Cys	8.5	-	CTG CTG	ā	CAG AAG TGG Gln Lys Trp 550	-	Prg.	-	AAG ATC Lys 11e	=	GTC CAG Val Gln			
1050 AGC TGC 6 Ser Cys 6	1140 CAC CCT (His Pro (1230 ACC TGC OThe Cys V	1320 CCT GAG Pro Glu P	1410 GAT GCC (Asp Ala 1	1500	GTC TTC (Val Phe	0	ATC TCC	1680	GAG CTG	1770	TTC AGC	1860	CGC TAC Arg Tyr			
cAG Gln	55 ti	CAG Gln	GAC Asp	ACA Thr		re C		CCC Arg		Ser		ATC 11e		GTC Val			
1040 TGC CCC Cys Pro	1130 ACG TGT The Cys	220 CTT TTG Leu Leu	310 CCC AGT Pro Ser	400 CCT CCC Pro Pro	1490	CTG GAC Leu Asp	1580	CGC CTC Arg Leu	1670	CGA CCC Arg Pro	1760	TTC CAA Phe Gln	1850	AAC TTT Asn Phe			
ACA TTG	CAA GTC	1220 GAT GAG CIT Asp Glu Leu	1310 TTG AAT CCC A	X 1400 TG GTG CCT al Val Pro		AGG CTA Arg Leu	7	ATG GAG	1	CCC AAG) i	-:	TCC CGG Ser Arg	15	SIn Car	
1030 ACG GCC A The Ala Ti	1120 GCC TGT G Ala Cys G 390	1210 c CTG G Leu A:	GTC ACC T	1390 1390 1 GGC CTG GTG G 1 Gly Leu Val V	1480	AGC Ser 510	1570	GAC ATG A	1660	G CAC C S Asp A 570	1750	AAA TAC ACA (1840	ATC Het	ရှိ ရှ 🗄	Lys 660	
A AG	55 22	1210 AAA ATC CTG (Lys Ile Leu /		86		TAC TGC Tyr Cys		5.5		CTC AAG GAC C		ĔĔ		CAA CGG Gln Arg	: 8	Asn Leu	
1020 ACC TGG The Tep	1110 TGT GCA Cys Ala	1200 CCA GGG Pro Gly	1290 GGA AAG Gly Lys	1380 GAG CCG Glu Pro	1470	GAT TTC Asp Phe	1560	TTT GTG Phe Val		ATC GGG	1740	GAG GTC	1830	GAG CCC Glu Pro	1920	His Ala	
one one /	AAT AGC 1 Asn Ser (TGC CCT C	GCC TCA C	TGC CAG (Cys Gln (TTG CAC Leu His A		AAG GCC Lys Ala		3.5		Ser	i	AGC CAG Ser Gln	1.0	Gly Pro	
010 AAG Lys	17. 17.	190 CAC His	8 E &	1370 GAA GCC TO Glu Ala C	1460	500	1550	CTC	0791	CAC GCC T	1730	TCC ACC	1820	P) P	1910	Gly 11e G	
CAT GGC H1s Gly	9 ₹	CAT GCC His Ala	형료	TCT GA/ Cys Glu		GAA CCG Glu Pro		GAA GTG CTG Glu Val Leu		GGC TCC Gly Ser		GTG GCC 1	1	CTG ATG Leu Het	1	Val.	
1000 GCC CAG Ala Gln 350	1090 G TGT GAG TGG u Cys Glu Trp 380	1180 GGC TGC (Gly Cys H	1270 GCT GGC CGG CAIA AIB A	1360 CTC ACC TGT (Leu Thr Cys (1450	ATC TCG Ile Ser 500	1540	GAG TTT Glu Phe 530	1630	CAC GAC H1s Asp 560	1720	25.06	1810	CTG CTC Leu Leu	1900	11e Pro 650	
Cys A	6A6 T	cAG G	val A	AAC C		GAC A		Ala		TAC		occ d		ACC TA	: 6	Val	
MVVF4	AHVUF4	ZHVUF4	MVVF4	MVVF4	Called	ZHANK ZHANET	63000	NHVWF4	130.4	MAUPA	FAUVE	MVVF4	AHVWF3	NHVUF4	AHVUF3		
		^	^	~	7	4 7	-		7	7.7	1	₹	Z	₹	7.7	•	
96	TC TGC -eu Cys 180		0		360	AGT					630	CAC Glu					990 AC GTG is Val
	TGC CTC Cys Leu 18	CCT GGA GAA	270	TCC ACG ATC Ser Thr 11e	360	TGC GGC AGT Cys Gly Ser	450	CTG GTG GAG Leu Val Glu	540	TCT GGC GGG Ser Gly Arg		TAC CAG GAG Tyr Gln Glu	720 CCT GTC GAC	Pro Val Asp	810 AAC AAC ATC Asn Asn Ile	900 GAG CCA TAT Glu Pro Tyr	CC CA
	: TCT GGC TGC : Ser Gly Cys 170	TAT GCC CCT GGA GAA Tyr Ala Pro Gly Glu	260 270	ACG TGC TCC ACG ATC Thr Cys Ser Thr 11e	350 360	GAT TAC TGC GGC AGT Asp Tyr Cys Gly Ser	440 450	ACC ATC CTG GTG GAG Thr Ile Leu Val Glu	530 540	CTG CAG TCT GGC CGG Val Glu Ser Gly Arg		CAG ACA TAC CAG GAG Gln Thr Tyr Gln Glu	710 720 GAA GAC CCT GTG GAC	Glu Asp Pro Val Asp	800 TGC CAT AAC AAC ATC Cys His Asn Asn Ile	GAC CCC GAG CCA TAT Asp Pro Glu Pro Tyr	980 GCC TAT GCC Ala Tyr Ala
08	tor ore ter eac rec bys Val Ser Gly Cys 170	TAT GCC CCT GGA GAA Tyr Ala Pro Gly Glu	260 270	AT GCC ACG TGC TCC ACG ATC ASP Ala Thr Cys Ser Thr Ile	350 360	GAT TAC TGC GGC AGT Asp Tyr Cys Gly Ser	440 450	TC ACC ATC CTG GTG GAG al Thr Ile Leu Val Glu	530 540	CTG CAG TCT GGC CGG Val Glu Ser Gly Arg	630	CTG AAG CAG ACA TAC CAG GAG Leu Lys Gln Thr Tyr Gln Glu	710 720 TG GAG GAA GAC CCT GTG GAC	ilu Glu Asp Pro Val Asp	800 CC TGC CAT AAC AAC ATC hr Cys His Asn Asn Ile	890 900 GTG GAC CCC GAG CCA TAT Val ASP Pro Clu Pro Tyr	980 GCC TAT GCC Ala Tyr Ala
08	tor ore ter eac rec bys Val Ser Gly Cys 170	TAT GCC CCT GGA GAA Tyr Ala Pro Gly Glu	260 270	AT GCC ACG TGC TCC ACG ATC ASP Ala Thr Cys Ser Thr Ile	140 350 360	CTG GTG CAG GAT TAC TGC GCC AGT Leu Val Gln Asp Tyr Cys Gly Ser 130	440 450	TC ACC ATC CTG GTG GAG al Thr Ile Leu Val Glu	530 540	CTG CAG TCT GGC CGG Val Glu Ser Gly Arg	630	GTC CTG AAG CAG ACA TAC CAG GAG Val Leu Lys Gln Thr Tyr Gln Glu 220	710 720 TG GAG GAA GAC CCT GTG GAC	ilu Glu Asp Pro Val Asp	800 CC TGC CAT AAC AAC ATC hr Cys His Asn Asn Ile	890 900 GTG GAC CCC GAG CCA TAT Val ASP Pro Clu Pro Tyr	980 GCC TAT GCC Ala Tyr Ala
08	AGC ATG GGC TGT GTC TCT GGC TGC Ser Net GJy Cys Val Ser GJy Cys 40 160 170	CAT CAG GGC AAG GAG TAT GGC CCT GGA GAA HIS GIN GIY LIJS GIU TYT AIR Pro GIY GIU	260 270	CAT GTG TGT GAT GCC AGG TGC TCC AGG ATC HIS Val Gys Asp Ala Thr Cys Ser Thr 11e	140 350 360	CTG GTG CAG GAT TAC TGC GCC AGT Leu Val Gln Asp Tyr Cys Gly Ser 130	430 440 450	TGC AAG AAA CGG GTC ACC ATC CTG GTG GAG Ays Lys Lys Lys Arg Val Thr 11e Leu Val GLu 160	530 540	CTG CAG TCT GGC CGG Val Glu Ser Gly Arg	620 630	TOC GTG GTC CTG AAG CAG ACA TAC CAG GAG Ser Val Val Leu Lys Gln Thr Tyr Gln Glu 220	700 710 720 AAC CTC CAA GTG GAG GAA CAC CCT GTG GAC	isn Leu Gln Val Glu Glu Asp Pro Val Asp 250	790 810 FICA TCC CCT GCC ACC TGC CAT AAC AAC ATC Ser Pro Ala Thr Cys His Asn Asn Ile 280	890 900 GTG GAC CCC GAG CCA TAT Val ASP Pro Clu Pro Tyr	980 GCC TAT GCC Ala Tyr Ala
60 70 80	TGC ATG AGC ATG GGC TGT CTCT GGC TGC Cys Net Ser Net GJy Cys Val Ser GJy Cys 40 150 160 170	CAT CAG GGC AAG GAG TAT GGC CCT GGA GAA HIS GIN GIY LIJS GIU TYT AIR Pro GIY GIU	240 250 260 270	CAT GTG TGT GAT GCC AGG TGC TCC AGG ATC HIS Val Gys Asp Ala Thr Cys Ser Thr 11e	140 350 360	CTG GTG CAG GAT TAC TGC GCC AGT Leu Val Gln Asp Tyr Cys Gly Ser 130	430 440 450	TGC AAG AAA CGG GTC ACC ATC CTG GTG GAG Ays Lys Lys Lys Arg Val Thr 11e Leu Val GLu 160	510 520 530 540	AND GAR ACT CAC TIT GAG GTG GTG GAG TCT GGC GGG GAG AND GAL Thr His Phe Glu Val Val Glu Ser Gly Arg	620 630	AGC ATC TCC GTG GTC CTG AAG CAG ACA TAC CAG GAG Ser lie Ser Val Val Leu Lys Gln Thr Tyr Gln Glu 220	690 700 700 710 720 AGC AGC AAC CTC GAA GTC GAA GAA GAC CCT GTC GAC	isn Leu Gln Val Glu Glu Asp Pro Val Asp 250	790 810 FICA TCC CCT GCC ACC TGC CAT AAC AAC ATC Ser Pro Ala Thr Cys His Asn Asn Ile 280	870 880 890 900 CAG CAC CAG CCA TAT CLIN ASP FOR CLIN CEN CCA TAT CLIN ASP FOR CLIN FOR TAT SIN ASP FOR CLIN FOR TYPE TAT SIN ASP FOR CLIN FOR TAT SING TAT	980 GCC TAT GCC Ala Tyr Ala
08 07 09	CTG GAC TYG AGC ATG GAC TGT GTC TCT GAC TGC Leu Glu Cys Net Ser Net Gly Cys Val Ser Gly Cys Val Cys Va	TOT CCC TGC TTC CAT CAG GGC AAG GAG TAT GCC CCT GGA GAA GAG TAT GCC TGC GGA GAA GAG TAT GCC CCT GGA GAA GAG TAT GCC CCT GGA GAA GAG FTC CT GGA GAA GAG FTC GGA GAG FTC GGA GAA GAA GAA GAA GAA GAA GAA GAA GA	240 250 260 270	TICE ACA GAC CAT GTG TGT GAT GCC ACG TGC TGC ACG ATC GYS Thr Asp His Val Gys Asp Ala Thr Cys Ser Thr 11e	140 350 360	GGG GAG TGC CAG TAC GTT CTG GTG CAG GAT TAC TGC GGC AGT GIG GIU Cys Gin Tyr Val Leu Val Gin Asp Tyr Cys Gly Ser 130	420 430 440 450	CCC TCA GTG AAA TGC AAA GGG GTC ACC ATC GTG GTG GAG Pro Ser Val Lys Cys Lys Lys Arg Val Thr Ile Leu Val Glu 160	510 520 530 540	ATG ANG GAT GAG ACT CAC TIT GAG GTG GTG GAG TCT GGC GGG Het Lys Asp Glu Thr Bis Phe Glu Val Val Glu Ser Gly Arg 190	610 620 630	CTG AGC ATC TCC GTG CTC CTG AAG CAG ACA TAC CAG GAG Leu Ser Ile Ser Val Val Leu Lys Gln Thr Tyr Gln Glu 220	690 700 710 720 CTC AC CAA GTG GAG GAA GAC CCT GTG GAC	Leu Thr Ser Ser Asn Leu Gln Val Glu Glu Asp Pro Val Asp	780 790 810 CCT CTG GAC TCA TCC CCT CAC CAC AAC ATC Pro Leu Asp Ser Ser Pro Ala Thr Cys His Asn Asn Ile 280	870 880 890 900 TTC CAG CAC TGC AAC AAG CTG GTG CAC CGC GAG CCA TAT Phe Cln Asp Cys Asu Iys Leu Val Asp Pro Clu Pro Tyr 310	980 GCC TAT GCC Ala Tyr Ala
08 07 09	CTG GAC TYG AGC ATG GAC TGT GTC TCT GAC TGC Leu Glu Cys Net Ser Net Gly Cys Val Ser Gly Cys Val Cys Va	TOT CCC TGC TTC CAT CAG GGC AAG GAG TAT GCC CCT GGA GAA GAG TAT GCC TGC GGA GAA GAG TAT GCC CCT GGA GAA GAG TAT GCC CCT GGA GAA GAG FTC CT GGA GAA GAG FTC GGA GAG FTC GGA GAA GAA GAA GAA GAA GAA GAA GAA GA	240 250 260 270	TICE ACA GAC CAT GTG TGT GAT GCC ACG TGC TGC ACG ATC GYS Thr Asp His Val Gys Asp Ala Thr Cys Ser Thr 11e	20 330 340 350 360	CC GGG GAG TGC CAG TAC GTT CTG GTG CAG GAT TAC TGC GGC AGT TAC GIV Cys GIn Tyr Val Leu Val Gin Asp Tyr Cys Gly Ser 130	110 420 430 440 450	ACC CC TCA GTG AAA TGC AAG AAA GGG GTC ACC ATC CTG GTG GAC ATS Pro Ser Val Lys Cys Lys Lys Arg Val Thr Ile Leu Val Glu 160	500 510 520 530 540	CCC ATG AAG GAT GAG ACT CAC TIT GAG GTG GTG GAG TCT GGC GGG PP MEL Lys Asp Glu Thr Bits Phe Glu Val Val Glu Ser Gly Arg	610 620 630	GGC CAC CTG AGC ATC TCC GTG GTC CTG AAG CAG AGA TAC CAG GAG Arg His Leu Ser Ile Ser Yal Val Leu Lys Gin Thr Tyr Gin Giu 220	690 700 710 720 CTC AC CAA GTG GAG GAA GAC CCT GTG GAC	sp Leu Thr Ser Ser Asn Leu Gln Val Glu Glu Asp Pro Val Asp	70 780 810 810 AA GTG CCT CTG GAC TCA CCT GCC ACC TCC CAT AAC AAC ATC ys Val Pro Leu Asp Ser Ser Pro Ala Thr Cys His Asn Asn Ile 280	860 870 880 890 900 GAC CTC CAG GAC CTC CAG CCC TAT Asp Val Pre Glin Asp Cys Am 159 Leu Val Asp Pro Glu Pro Tyr 310	980 GCC TAT GCC Ala Tyr Ala
08 07 09	CTG GAC TYG AGC ATG GAC TGT GTC TCT GAC TGC Leu Glu Cys Net Ser Net Gly Cys Val Ser Gly Cys Val Cys Va	TOT CCC TGC TTC CAT CAG GGC AAG GAG TAT GCC CCT GGA GAA GAG TAT GCC TGC GGA GAA GAG TAT GCC CCT GGA GAA GAG TAT GCC CCT GGA GAA GAG FTC CT GGA GAA GAG FTC GGA GAG FTC GGA GAA GAA GAA GAA GAA GAA GAA GAA GA	240 250 260 270	TICE ACA GAC CAT GTG TGT GAT GCC ACG TGC TGC ACG ATC GYS Thr Asp His Val Gys Asp Ala Thr Cys Ser Thr 11e	20 330 340 350 360	CC GGG GAG TGC CAG TAC GTT CTG GTG CAG GAT TAC TGC GGC AGT TAC GIV Cys GIn Tyr Val Leu Val Gin Asp Tyr Cys Gly Ser 130	110 420 430 440 450	ACC CC TCA GTG AAA TGC AAG AAA GGG GTC ACC ATC CTG GTG GAC ATS Pro Ser Val Lys Cys Lys Lys Arg Val Thr Ile Leu Val Glu 160	500 510 520 530 540	CCC ATG AAG GAT GAG ACT CAC TIT GAG GTG GTG GAG TCT GGC GGG PP MEL Lys Asp Glu Thr Bits Phe Glu Val Val Glu Ser Gly Arg	590 600 610 620 630	AND CRC CAC CTG AGG ATC TCG CTG CTG AAG CAG AGA TAC CAG GAG AGA TAC CAG GAG INP ATB HIS Leu Ser Ile Ser Val Val Leu Lys Gin Thr Tyr Gin Glu 220	680 690 700 710 720 AAC AAC CTC CAA GTG GAG GAA GAA GAC CCT GTG GAC	Asn Asn Leu Thr Ser Ser Asn Leu Gln Val Glu Glu Asp Pro Val Asp	70 780 810 810 AA GTG CCT CTG GAC TCA CCT GCC ACC TCC CAT AAC AAC ATC ys Val Pro Leu Asp Ser Ser Pro Ala Thr Cys His Asn Asn Ile 280	860 870 880 890 900 900 FG CAC CAC CAC CAC CAC TAT FEE ASP Val Phe Cln Asp Cys Am Lys Leu Val Asp Pro Clu Pro Tyr 310	980 GCC TAT GCC Ala Tyr Ala
08 07 09	CTG GAC TYG AGC ATG GAC TGT GTC TCT GAC TGC Leu Glu Cys Net Ser Net Gly Cys Val Ser Gly Cys Val Cys Va	TOT CCC TGC TTC CAT CAG GGC AAG GAG TAT GCC CCT GGA GAA GAG TAT GCC TGC GGA GAA GAG TAT GCC CCT GGA GAA GAG TAT GCC CCT GGA GAA GAG FTC CT GGA GAA GAG FTC GGA GAG FTC GGA GAA GAA GAA GAA GAA GAA GAA GAA GA	240 250 260 270	TICE ACA GAC CAT GTG TGT GAT GCC ACG TGC TGC ACG ATC GYS Thr Asp His Val Gys Asp Ala Thr Cys Ser Thr 11e	20 330 340 350 360	CC GGG GAG TGC CAG TAC GTT CTG GTG CAG GAT TAC TGC GGC AGT TAC GIV Cys GIn Tyr Val Leu Val Gin Asp Tyr Cys Gly Ser 130	110 420 430 440 450	ACC CC TCA GTG AAA TGC AAG AAA GGG GTC ACC ATC CTG GTG GAC ATS Pro Ser Val Lys Cys Lys Lys Arg Val Thr Ile Leu Val Glu 160	500 510 520 530 540	CCC ATG AAG GAT GAG ACT CAC TIT GAG GTG GTG GAG TCT GGC GGG PP MEL Lys Asp Glu Thr Bits Phe Glu Val Val Glu Ser Gly Arg	600 610 620 630	AND CRC CAC CTG AGG ATC TCG CTG CTG AAG CAG AGA TAC CAG GAG AGA TAC CAG GAG INP ATB HIS Leu Ser Ile Ser Val Val Leu Lys Gin Thr Tyr Gin Glu 220	680 690 700 710 720 AAC AAC CTC CAA GTG GAG GAA GAA GAC CCT GTG GAC	Asn Asn Leu Thr Ser Ser Asn Leu Gln Val Glu Glu Asp Pro Val Asp	70 780 810 810 AA GTG CCT CTG GAC TCA CCT GCC ACC TCC CAT AAC AAC ATC ys Val Pro Leu Asp Ser Ser Pro Ala Thr Cys His Asn Asn Ile 280	860 870 880 890 900 900 FG CAC CAC CAC CAC CAC TAT FEE ASP Val Phe Cln Asp Cys Am Lys Leu Val Asp Pro Clu Pro Tyr 310	980 GCC TAT GCC Ala Tyr Ala
08 07 09	CTG GAC TYG AGC ATG GAC TGT GTC TCT GAC TGC Leu Glu Cys Net Ser Net Gly Cys Val Ser Gly Cys Val Cys Va	TOT CCC TGC TTC CAT CAG GGC AAG GAG TAT GCC CCT GGA GAA GAG TAT GCC TGC GGA GAA GAG TAT GCC CCT GGA GAA GAG TAT GCC CCT GGA GAA GAG FTC CT GGA GAA GAG FTC GGA GAG FTC GGA GAA GAG FTC GGA GAG FTC GGA GAA GAG FTC GGA GAG FTC	240 250 260 270	TICE ACA GAC CAT GTG TGT GAT GCC ACG TGC TGC ACG ATC GYS Thr Asp His Val Gys Asp Ala Thr Cys Ser Thr 11e	310 320 330 340 350 360	CTC AAA TAC CTG TTC CCC GGG GAG TGC CAG TAC GTG GGG GAT TAC TGC GGG ACT TAC TGC GGG ACT TAC TGC GGG ACT TAC TGC GGG GAT TAC TGC GGG ACT TAC TGC GGG GAT TAC TGC GGG ACT TGC ACT TG	110 420 430 440 450	ACC CC TCA GTG AAA TGC AAG AAA GGG GTC ACC ATC CTG GTG GAC ATS Pro Ser Val Lys Cys Lys Lys Arg Val Thr Ile Leu Val Glu 160	500 510 520 530 540	CCC ATG AAG GAT GAG ACT CAC TIT GAG GTG GTG GAG TCT GGC GGG PP MEL Lys Asp Glu Thr Bits Phe Glu Val Val Glu Ser Gly Arg	580 590 600 610 620 630	TOC GTG GTG TGG GAC GGC GAC GTG AGC ATC TGC GTG GTG CTG AAG GAG AGA TAC CAG GAG Ser Val Val Trp Asp Arg His Leu Ser Ile Ser Val Val Leu Lys Gin Thr Tyr Gin Glu 210	680 690 700 710 720 AAC AAC CTC CAA GTG GAG GAA GAA GAC CCT GTG GAC	21y Ile Gin Asn Asn Asp Lev Thr Ser Ser Asn Lev Gin Val Giu Giu Asp Pro Val Asp 240	100 TOT GCT GCT ACA ACA GTG CCT CTG GAC TCA TCC CCT GCC ACC TCC CAT AAA GAC ATC ATC CCT GCT CCT CCT CCT CCT CCT CCT CCT TCC ATC ACT TCC CT CCT TCC ATC TCC ATC AT	850 860 900 800 800 900 800 890 900 900 841 CTT ACC ANT CAC CAC CAC CAC CCC CAC CCA TAT ARE ILE LEAU THY SER ASP VAI PIPE CIN ASP CYS ASP LIYS LEAU VAI ASP Pro CIU Pro Tyr 300	980 GCC TAT GCC Ala Tyr Ala
08 07 09	CTG GAC TYG AGC ATG GAC TGT GTC TCT GAC TGC Leu Glu Cys Net Ser Net Gly Cys Val Ser Gly Cys Val Cys Va	TOT CCC TGC TTC CAT CAG GGC AAG GAG TAT GCC CCT GGA GAA GAG TAT GCC TGC GGA GAA GAG TAT GCC CCT GGA GAA GAG TAT GCC CCT GGA GAA GAG FTC CT GGA GAA GAG FTC GGA GAG FTC GGA GAA GAG FTC GGA GAG FTC GGA GAA GAG FTC GGA GAG FTC	240 250 260 270	TICE ACA GAC CAT GTG TGT GAT GCC ACG TGC TGC ACG ATC GYS Thr Asp His Val Gys Asp Ala Thr Cys Ser Thr 11e	310 320 330 340 350 360	CTC AAA TAC CTG TTC CCC GGG GAG TGC CAG TAC GTG GGG GAT TAC TGC GGG ACT TAC TGC GGG ACT TAC TGC GGG ACT TAC TGC GGG GAT TAC TGC GGG ACT TAC TGC GGG GAT TAC TGC GGG ACT TGC ACT TG	110 420 430 440 450	ACC CC TCA GTG AAA TGC AAG AAA GGG GTC ACC ATC CTG GTG GAC ATS Pro Ser Val Lys Cys Lys Lys Arg Val Thr Ile Leu Val Glu 160	500 510 520 530 540	CCC ATG AAG GAT GAG ACT CAC TIT GAG GTG GTG GAG TCT GGC GGG PP MEL Lys Asp Glu Thr Bits Phe Glu Val Val Glu Ser Gly Arg	580 590 600 610 620 630	GCC CTC TCC CTG GTC TGG GAC GCC CAC CTG AGC ATC TCC GTG GTC CTG AAG CAG ACA TAC CAG GAG ALA Leu Ser Val Val Leu Lys Gln Thr Tyr Gln Glu Ala Leu Ser Val Val CLn Thr Tyr Gln Glu 220	680 690 700 710 720 AAC AAC CTC CAA GTG GAG GAA GAA GAC CCT GTG GAC	21y Ile Gin Asn Asn Asp Lev Thr Ser Ser Asn Lev Gin Val Giu Giu Asp Pro Val Asp 240	100 TOT GCT GCT ACA ACA GTG CCT CTG GAC TCA TCC CCT GCC ACC TCC CAT AAA GAC ATC ATC CCT GCT CCT CCT CCT CCT CCT CCT CCT TCC ATC ACT TCC CT CCT TCC ATC TCC ATC AT	840 850 860 990 990 TCC TCC AAC AAC CTC CTC CAC CCC CAC CCC CAC CCA TAT Set Cys Arg I le Leu Thr Set Asp Val Phe Clin Asp Cys Arg I le Leu Thr Set Asp Val Phe Clin Asp Cys Arg I le Leu Thr Set Cys Arg I le Unit Asp Cys Arg I le Unit Asp Pro Clu Pro Tyr 310	980 GCC TAT GCC Ala Tyr Ala
08 07 09	CTG GAC TYG AGC ATG GAC TGT GTC TCT GAC TGC Leu Glu Cys Net Ser Net Gly Cys Val Ser Gly Cys Val Cys Va	TOT CCC TGC TTC CAT CAG GGC AAG GAG TAT GCC CCT GGA GAA GAG TAT GCC TGC GGA GAA GAG TAT GCC CCT GGA GAA GAG TAT GCC CCT GGA GAA GAG FTC CT GGA GAA GAG FTC GGA GAG FTC GGA GAA GAG FTC GGA GAG FTC GGA GAA GAG FTC GGA GAG FTC	210 220 230 240 250 260 270 A A A A A A A A A A A A A A A A A A A	NCT TOT CIT THE GAG COGG AMG TOG GAG TOC ACA CAC CAT CIT CIT CAC ACG TOC TOC ACG ATC TOC TOT TOT TOT TOT TOT TOT TOT TOT T	310 320 330 340 350 360	CTC AAA TAC CTG TTC CCC GGG GAG TGC CAG TAC GTG GGG GAT TAC TGC GGG ACT TAC TGC GGG ACT TAC TGC GGG ACT TAC TGC GGG GAT TAC TGC GGG ACT TAC TGC GGG GAT TAC TGC GGG ACT TGC ACT TG	110 420 430 440 450	ACC CC TCA GTG AAA TGC AAG AAA GGG GTC ACC ATC CTG GTG GAC ATS Pro Ser Val Lys Cys Lys Lys Arg Val Thr Ile Leu Val Glu 160	500 510 520 530 540	CCC ATG AAG GAT GAG ACT CAC TIT GAG GTG GTG GAG TCT GGC GGG PP MEL Lys Asp Glu Thr Bits Phe Glu Val Val Glu Ser Gly Arg	570 580 590 600 610 620 630	GGC AAA GGC CTC TOC GTG CTC TOG GAC GGC CAC CTG AGC ATC TOC GTG CTC CTG AAA CAA TAC CAG GAG GJA AAA GGC TT TOC TTG AAA AAA AAAA AAAA AAAA AAAA AAAA A	660 670 680 690 700 700 710 720 336 AAT TIT GAT GAC CTC CAA CTC CAA CTC GAA GAC CCT CAC GAC CAC CAC CAC CAC CAC CAC CAC CAC	Asn Phe Asp GIy Lie Cin Asn Asn Asp Leu Thr Ser Ser Asn Leu Cin Val Ciu Giu Asp Pro Val Asp	ANCE THE GET GET GAT GAC ACC ACA AAA GTE CET CTG GAC TEA TOC CET CCC ACC TEC CAT AAAC AAAC ATC SEE SEE SEE GIN CYS ALA AFB THIT ATE LIYS VAI PTO Leu ASP SEE SEE PTO ALIA THIT CYS HIS ASN ASN IITE 270	840 850 860 990 990 TCC TCC ACT ALA ATC CLT ACC ACT TCC AAC AAC CTC GAC CCC GAC CCA TAT Set Set Cys Arg Ile Leu Tht Set Asp Val Phe Cln Asp Cys Ann Iys Leu Val Asp Pro Clu Pro Tyr 300	980 GCC TAT GCC Ala Tyr Ala
20 30 40 50 60 70 80 CG C C C Ala	CTC GAG TOTA ÄGT AAA AGG TOC CAG AAC TAT GAG CTG GAG TOC ATG AGG ATG GAC TOT CTC CTC TCC TGC Lev Glu Gys Met Ser Met Gly Gys Val Ser Gly Gys 100 120 130 140 150 150 150 170 170 170 170 170 170 170 170 170 17	TOT CCC TGC TTC CAT CAG GGC AAG GAG TAT GCC CCT GGA GAA GAG TAT GCC TGC GGA GAA GAG TAT GCC CCT GGA GAA GAG TAT GCC CCT GGA GAA GAG FTC CT GGA GAA GAG FTC GGA GAG FTC GGA GAA GAG FTC GGA GAG FTC GGA GAA GAG FTC GGA GAG FTC	200 210 220 230 240 250 260 270 A A A A A A A A A A A A A A A A A A A	CGC TGC AAC TGT GTC TGT GAG GGG AAC TGG TGC ACA GAC CAT GTC TGT GAT GCC ACG TGC ACG GTC GLG Cys Asn The Cys Val Cys Gat Lys Trp Asp Cys The Asp His Val Cys Asp His Cys Set The Tip Asp Cys The Asp His Val Cys Asp His Val Cys Cys The Asp Cys The Asp His Val Cys Asp His Val Cys Cys The Tip Asp Cys The Asp His Val Cys Asp His Val Cys Cys The Tip Asp Cys The Asp His Val Cys Cys Cys The Tip Asp Cys The Asp His Val Cys Cys Cys The Cys	310 320 330 340 350 360	CTC AAA TAC CTG TTC CCC GGG GAG TGC CAG TAC GTG GGG GAT TAC TGC GGG ACT TAC TGC GGG ACT TAC TGC GGG ACT TAC TGC GGG GAT TAC TGC GGG ACT TAC TGC GGG GAT TAC TGC GGG ACT TGC ACT TG	110 420 430 440 450	TITY GGG ATC CTA GTG GGG ANY AMG GGA TGG AGC CAC CCC TGA GTG AMA TGC AMG AMA GGG GTG ACC ATC CTG GTG GAG Phe Atg Ile Leu Val Gly Asn Lys Gys Set His Pro Set Val Lys Cys Lys Lys Arg Wal Thr Ile Leu Val Glu 160	470 480 490 500 510 520 530 540	A. CTC TIT GAC GGG GAC GTG ANT GTG ANG AGG CGG ATG AAG GAT GAC TO CAT TIT GAG GTG GTG GAG TIT GGC GGG GAG ILL Leu Phe Asp Gly Glu Val Asn Asp Asp Pro Met Lys Asp Glu The His Phe Glu Val Val Glu Ser Gly Arg	580 590 600 610 620 630	CTG CTG GGG AAA GGC CTC TCC CTG GTC TGG GAC GGC CAC CTG AGC ATC TCC GTG CTC CTC AAA CAA AAA TAC CAG GAG Leu Leu GLY Lys Ala Leu Ser Val Val TTP ASP ATG HIS Leu Ser Ile Ser Val Val Leu Lys Gln Thr Tyr Gln Glu 220	650 660 670 680 690 700 700 710 720 CTC TGT GGG ANT TIT GAT GGC ATC CAC ACT CAC ACT CAC ACT CAC ACC CAC CA	Leu Cys Gly Asn Phe Asp Gly Ile Gln Asn Asn Leu Thr Ser Ser Asn Leu Gln Val Glu Glu Asp Pro Val Asp 240	THE AAM GITE AGE OF CHE GET GET GAC AGE WEAR AAM GITE CET GAC TEA THE CEC CET GAC CET CEC CAT AAC AAC ATC TE LYS VAI SET SET CIT GAY AND ALC ATC TE LYS VAI SET SET CO ALLA THE CYS HIS ASM ASM INE TIPE LYS VAI SET SET ASM ASM INC.	830 840 850 950 900 AT CT ACA AT CAA CTC TTC CAC CAC TCC AAC AA	980 GCC TAT GCC Ala Tyr Ala
10 20 30 40 50 60 70 80 C- C Ala	ocy can eac cric dao tor AGE AMA Aos toc cad Mc Tar Cad cric dat toc ATO Ato Ato Ato Ato Car cric cric roc toc roc at Ala Glu Gly Leu Glu Gys Thr Lys Thr Cys Cla Ann Tyr Asp Leu Glu Gys Het Ser Het Gly Gys Val Ser Gly Gys 20 10 110 120 130 140 150 150 160 170	TOT CCC TGC TTC CAT CAG GGC AAG GAG TAT GCC CCT GGA GAA GAG TAT GCC TGC GGA GAA GAG TAT GCC CCT GGA GAA GAG TAT GCC CCT GGA GAA GAG FTC CT GGA GAA GAG FTC GGA GAG FTC GGA GAA GAG FTC GGA GAG FTC GGA GAA GAG FTC GGA GAG FTC	200 210 220 230 240 250 260 270 A A A A A A A A A A A A A A A A A A A	NIT GGC TGC AAG TGT TGT GAG GGG AAG TGG GAG TGC ACA GAC CAT GTG TGT GAT GCC ACG TGC ACG ATC TGT GAT GG TGC ACG ATC TGT GAT GG TGC ACG AGG TGC ACG ATC TGT GAT GG TGC ACG AGG TGC ACG ATC TGT GAT GGG AGG TGC ACG ACG ACG ACG ACG ACG ACG ACG ACG A	- 290 300 310 320 330 340 350 360 	DEC CAC TAC CTC ACC TTC CAC CAC CAC AND TAC CTC TTC CCC CAG CAC CAC CAC CAC CAC CAC CAC CA	380 390 400 410 420 430 440 450	LOC TIT GGG ATC GTG GGG ANY AMG GGA TGG AGC CAC CCC TGA GTG ANA TGC AMG ANG GGG GTG ACC ATC GTG GAG Thr. Phe Arg Ile Leu Val Gly Asn Lys Gys Ser His Pro Ser Val Lys Cys Lys Lys Arg Wal Thr Ile Leu Val Glu Thr Ile Leu Val Glu Thr Ile Cac Ang Arg Wal Thr Ile Leu Val Glu Thr Ile Cac Ang Arg Wal Thr Ile Leu Val Glu Thr Ile Cac Ang Arg Wal Thr Ile Leu Val Glu Thr Ile Cac Ang Arg Wal Thr Ile Leu Val Glu Thr Ile Cac Ang Arg Wal Thr Ile Leu Val Glu Thr Ile Cac Ang Arg Wal Thr Ile Leu Val Glu Thr Ile Cac Ang Arg Wal Thr Ile Leu Val Glu Thr Ile Cac Ang Arg Wal Thr Ile Leu Val Glu Thr Ile Cac Ang Arg Wal Thr Ile Leu Val Glu Thr Ile Cac Ang Arg Wal Thr Ile C	470 480 490 500 510 520 530 540	A. CTC TIT GAC GGG GAC GTG ANT GTG ANG AGG CGG ATG AAG GAT GAC TO CAT TIT GAG GTG GTG GAG TIT GGC GGG GAG ILL Leu Phe Asp Gly Glu Val Asn Asp Asp Pro Met Lys Asp Glu The His Phe Glu Val Val Glu Ser Gly Arg	560 570 580 590 600 610 620 630	ATT CTG CTG CTG AAA GGC CTC TCG CTG GTC TGG GAC CGC CAC CTG AGC ATC CCG CTG CTG CAC CAA GAC AAA CAC TAC CAG GAG 11s Leu Leu Leu Gly Lys Ala Leu Ser Val WI TTP ASP ATG HIS Leu Ser Ile Ser Val 200 Leu Lys Gln Thr Tyr Gln Glu 210	650 660 670 680 690 700 700 710 720 GCC CTC TOT GCC ANT TIT CAT GCC ATC CAC ACT CAC ACT CAC ACT CAC ACC CAC CA	Gly Leu Cys Gly Asn Phe Asp Gly 11e Gln Asn Asn Asp Leu Thr Ser Ser Asn Leu Gln Val Glu Glu Asp Pro Val Asp 240	TOC TOG AND GTO AGC TOT GOT GOT GOT GOT AND AND GTO CCT CTG GAV TOCA TOC GCC ACC TOC CAT AND	830 840 850 860 890 900 000 AT COLOR TAT THE MET WELL MAP SET	980 GCC TAT GCC Ala Tyr Ala
10 20 30 40 50 60 70 80	CCC CCT CAA CCC CAC TCT AGE AAA ACS TCC CAG AAC TAT CAAC CTC CAG CTC ATC ACC ATC ACC TCT CTC TCT CCC TCC ACC A	CCC GCC ATG GTC CGC CAT GAG ARA TOT GTG GCC GTG GAA AGC TGT CCC TGC TTT CAT CAG GCC AAG GAG TAT GCC CCT GGA GAA Pro GLJ Met Val Arg His Glu Asn Arg Cys Val Ala Leu Glu Arg Cys Pro Cys Phe His Glu Gly Lys Glu Tyr Ala Pro Gly Glu 50	200 210 C 220 230 A 240 250 260 270 A 448 A 240 250 270 270 270 2 270 2 2 2 2 2 2 2 2 2 2	NIT GGC TGC AAG TGT TGT GAG GGG AAG TGG GAG TGC ACA GAC CAT GTG TGT GAT GCC ACG TGC ACG ATC TGT GAT GG TGC ACG ATC TGT GAT GG TGC ACG AGG TGC ACG ATC TGT GAT GG TGC ACG AGG TGC ACG ATC TGT GAT GGG AGG TGC ACG ACG ACG ACG ACG ACG ACG ACG ACG A	310 320 330 340 350 360	DEC CAC TAC CTC ACC TTC CAC CAC CAC AND TAC CTC TTC CCC CAG CAC CAC CAC CAC CAC CAC CAC CA	110 420 430 440 450	LOC TIT GGG ATC GTG GGG ANY AMG GGA TGG AGC CAC CCC TGA GTG ANA TGC AMG ANG GGG GTG ACC ATC GTG GAG Thr. Phe Arg Ile Leu Val Gly Asn Lys Gys Ser His Pro Ser Val Lys Cys Lys Lys Arg Wal Thr Ile Leu Val Glu Thr Ile Leu Val Glu Thr Ile Cac Ang Arg Wal Thr Ile Leu Val Glu Thr Ile Cac Ang Arg Wal Thr Ile Leu Val Glu Thr Ile Cac Ang Arg Wal Thr Ile Leu Val Glu Thr Ile Cac Ang Arg Wal Thr Ile Leu Val Glu Thr Ile Cac Ang Arg Wal Thr Ile Leu Val Glu Thr Ile Cac Ang Arg Wal Thr Ile Leu Val Glu Thr Ile Cac Ang Arg Wal Thr Ile Leu Val Glu Thr Ile Cac Ang Arg Wal Thr Ile Leu Val Glu Thr Ile Cac Ang Arg Wal Thr Ile Leu Val Glu Thr Ile Cac Ang Arg Wal Thr Ile C	460 470 480 490 500 510 520 530 540	GOA GAG ATT GAC CTO TIT CAC GOG GAG GTG AAT GTG AAG GCC ATG AAG GAT GAG ATT CAC TIT GAG GTG GTG GAG TOT GGC GGG GIG GAU THE GAG TOT GAG GAG TOT GAG GAG GAG GAG GAG GAG GAG GAG GAG GA	570 580 590 600 610 620 630	ATC ATT CTC CTC CTC GGC AAA GGC CTC TGC GTC GTC GGC GGC CTC GGC GTC GT	640 650 660 670 680 690 730 720 CTC ACC ACC ACC ACC ACC CTC CAA CTC CACC ACC	Val Cys Gly Leu Cys Gly Asn Phe Asp Gly 11e Gln Asn Asn Asp Leu Thr Ser Ser Asn Leu Gln Val Glu Asp Pro Val Asp 230 120 120 120 120 120 120 120 120 120 12	GGG ANC TOC TOC ANA OTT ACT TOC ACT TOT GGT AGG ACC AGA ANA OTT CCTT GAC TOCA TOC COT GCC ACC TOC CAT AND AAC ANC ATC GLY ASS SET TOP 195 VAI SET SET GIN Gys Ala ABP The Arg Lys Val Pro Leu Asp Set Set Pro Ala Thr Cys His Asm Asn Lite 260	820 830 840 850 860 870 880 880 990 900 AME CAG ACT OFF CAG	980 GCC TAT GCC Ala Tyr Ala
10 20 30 40 50 60 70 80	AC CTG CGG CAT GAA AGG TGC CAG TGT ÄGT AAA AGG TGC CAG AAC TAT CAAC CTG CAG CATC ATC ATC GCT TGT CTGT C	TOT CCC TGC TTC CAT CAG GGC AAG GAG TAT GCC CCT GGA GAA GAG TAT GCC TGC GGA GAA GAG TAT GCC CCT GGA GAA GAG TAT GCC CCT GGA GAA GAG FTC CT GGA GAA GAG FTC GGA GAG FTC GGA GAA GAG FTC GGA GAG FTC GGA GAA GAG FTC GGA GAG FTC	190 200 210 220 230 240 250 260 270 A A A A A A A A A A A A A A A A A A A	ATT GGC TOC ANC ACT TOT GTC TOT GAG GAC GGG ANG TGG GAT GTC ACA GAC CAT GTC TOT GAT GCC ACG TCC ACG ATC ITE GJY SAS THI CYS VAI CYS CAT A 90 TTP ASP GYS THI ASP HIS VAI CYS SET THI ITE	280 290 300 310 320 330 340 350 360	CTC AAA TAC CTG TTC CCC GGG GAG TGC CAG TAC GTG GGG GAT TAC TGC GGG ACT TAC TGC GGG ACT TAC TGC GGG ACT TAC TGC GGG GAT TAC TGC GGG ACT TAC TGC GGG GAT TAC TGC GGG ACT TGC ACT TG	370 380 390 400 410 420 430 440 450	TITY GGG ATC CTA GTG GGG ANY AMG GGA TGG AGC CAC CCC TGA GTG AMA TGC AMG AMA GGG GTG ACC ATC CTG GTG GAG Phe Atg Ile Leu Val Gly Asn Lys Gys Set His Pro Set Val Lys Cys Lys Lys Arg Wal Thr Ile Leu Val Glu 160	460 470 480 490 500 510 520 530 540	A. CTC TIT GAC GGG GAC GTG ANT GTG ANG AGG CGG ATG AAG GAT GAC TO CAT TIT GAG GTG GTG GAG TIT GGC GGG GAG ILL Leu Phe Asp Gly Glu Val Asn Asp Asp Pro Met Lys Asp Glu The His Phe Glu Val Val Glu Ser Gly Arg	550 560 570 580 590 600 610 620 630	ATT CTG CTG CTG AAA GGC CTC TCG CTG GTC TGG GAC CGC CAC CTG AGC ATC CCG CTG CTG CAC CAA GAC AAA CAC TAC CAG GAG 11s Leu Leu Leu Gly Lys Ala Leu Ser Val WI TTP ASP ATG HIS Leu Ser Ile Ser Val 200 Leu Lys Gln Thr Tyr Gln Glu 210	650 660 670 680 690 700 700 710 720 GCC CTC TOT GCC ANT TIT CAT GCC ATC CAC ACT CAC ACT CAC ACT CAC ACC CAC CA	Val Cys Gly Leu Cys Gly Asn Phe Asp Gly 11e Gln Asn Asn Asp Leu Thr Ser Ser Asn Leu Gln Val Glu Asp Pro Val Asp 230 120 120 120 120 120 120 120 120 120 12	TOC TOG AND GTO AGC TOT GOT GOT GOT GOT AND AND GTO CCT CTG GAV TOCA TOC GCC ACC TOC CAT AND	830 840 850 860 890 900 000 AT COLOR TAT THE MET WELL MAP SET	TAT GCC Tyr Ala

FIGURE 3: Nucleotide and translated amino acid sequence of human von Willebrand factor cDNA isolates. The sequence of AHvWF4 is shown in its entirety, and the nucleotide numbering begins with the first nucleotide of that isolate. The amino acid numbering assumes that the amino-terminal Ser of plasma von Willebrand factor is residue 1. Identity between regions of AHvWF1, AHvWF2, and AHvWF3 that overlap with AHvWF4 are indicated by dashes (---), and discrepancies are shown explicitly. Where the translated amino acid sequences of the cDNA isolates disagree, the amino

acid that agrees with the major sequence determined independently for the protein (Tiani et al., 1986) is underlined. Potential Asn-X-Thr/Ser glycosylation sites that are utilized in von Willebrand factor are indicated by closed triangles (*), and one potential site that is not utilized is indicated by an X. The location of a glycosylated Asn in the sequence Asn-Ser-Cys is shown by a closed circle (*).

3168 BIOCHEMISTRY SHELTON-INLOES ET AL.

DOMAIN B Triplication 1 5	10 15	20 25	30	35	
1533 CPTAKAPT	CGLCEVAR	LRQNADQCCI	PEYECVCDP	Ý	
15// CERGLOPT		LTNP-GBCRI LRKTOCCI	PNFTCACKK	ь	
DOMAIN D Duplication	10 15	20, 25,	30	35, 40,	45 50 55 60
79	V C R D R K W N	CTDHVCDAT	C S T I G M A H Y	LTPDGLKYL	F P G E C Q Y V L V Q D Y C G S N L T G S C S Y V L P Q N K
					<u> </u>
, ,	70 75 GNKGCSHP	80 85 S V K – - C K K R V		95, 100, I B L F D G – E V	N V K R P M K D E T H F E V - V E
1218 E Q D L E Y I L	HNGACS-P	AROGCHKS	I B V K H S A L S	VELESDMEV	T V N G R L V S V P Y V G G N H B
121 125	130 135	140 145	150 1	160	165 170 175 180
195 S G R Y I I L L 1277 V N V Y G A I M	L HEVRFNHL	G K A L S V V V D E	R H L S I S V V L O N N E F O L O L	K - QTYQEKV SPKTFASKT	C G L C G N P D G I Q N N D L T S Y G L C G I C D E N G A N D F M L
181 185					
, ,	, ,	200, 205 KVSSQCADTI	210, 2 R K V P L <u>D</u> S <u>S</u> P	15 220, A T C <u>H</u> N N <u>I</u> H K	, , , , , , ,
1336 R D G T V T T D	W K T L V Q E W	r v	<u>Q</u> R <u>P</u> G	Q T C Q P I L B B	Q C L V P D S S H C Q V L L L P L
241 245	250 255	260 265	270 2	75 280	285 290 295 300
305 F Q D C N K L V 1384 F A E C H K V L	DPEPYLD <u>V</u>	CIYDTCSCES	S <u>I</u> G D C A C F C Q V C	DTIAAYAH <u>V</u> EVIASYAHL	C A Q E G K V V T W R T A T L C P C R T N G V C V D W R T P D F C A
301	· [][ı., ⊓-⊢ .[⊓	· ~	- <u> </u>	
365 Q S C					
1438 M S C					
DOMAIN E Duplication					
1 5	10 15	20 25	30	35 40	45

25 CTKTCQNYDLECMSMGCVSGCLCPPGMVRHENRCVALBRCP-CPHQG
1453 CPRHCDGNVSSCGDHPS-EGCFCPPDKVMLEGSCVPBBBACTQCIGED

FIGURE 4: Alignment of repeated amino acid sequence domains in human von Willebrand factor. The amino acid numbering begins with the amino-terminal Ser of the plasma von Willebrand factor subunit. The alignments shown were obtained with the Align computer program (Dayhoff et al., 1983). The scoring matrix used was the mutation data matrix (MD+2), with a gap penalty of 6. Amino acids are shown by the single-letter code. Identical residues are shown enclosed in boxes, and conservative substitutions are underlined. In this figure, a conservative substitution is one with a positive value in the mutation data matrix (Dayhoff et al., 1983). Statistical parameters for these alignments are listed in Table I.

of λ HvWF3 (Sadler et al., 1985) and that reported by Lynch et al. (1985).

There are two potential N-glycosylation sites in the translated amino acid sequence of $\lambda HvWF4$ that lies in the gap between $\lambda HvWF1$ and $\lambda HvWF3$ (Figure 3). Thus, there are a total of 13 Asn-X-Thr/Ser sequences in the von Willebrand factor subunit. The first of these, at Asn-94, is abolished by the A to G transition at nucleotide 231 of $\lambda HvWF4$. The potential site at Asn-452 is not glycosylated, and Asn-384, in the unusual sequence Asn-Ser-Cys, is glycosylated in plasma von Willebrand factor. The actual sites of Asn-linked and Thr/Ser-linked glycosylation have been determined directly by protein sequencing (Titani et al., 1986).

Internal Homologies in von Willebrand Factor Protein Sequence. Five unrelated sequences are repeated within the von Willebrand factor subunit. Three of these (domains A-C) have been described previously (Sadler et al., 1985). Domain A contains 193-220 amino acids and is present in three tandem copies between residues 497 and 1111 (not shown). The sequence of λHvWF4 (Figure 3) completes the amino-terminal sequence of domain A1. Domain B contains 25-35 amino acids and is present in three copies between residues 1533 and

1636 (Figure 4). Further analysis has revealed a third copy (domain B2) between the previously reported pair. The statistical similarity of domain B2 to domains B1 and B3 is the least significant of all segment comparisons (Table I), but the length of domain B2 is the same as that of domain B3, many residues (particularly Cys) are shared by all three domains, and domain B2 is precisely centered between B1 and B3. Thus, we propose domain B2 as a faint but definite member of the domain B group of peptide segments. Domain C contains 116-119 amino acids and is duplicated between residues 1637 and 1899 (not shown).

In contrast to the virtually contiguous replication of domains A-C, two additional types of repeated domain can be identified within the sequence of $\lambda HvWF4$ or $\lambda HvWF3$ that are separated from their homologues by unrelated repeated domains. Domain D1 contains 289 amino acids between residues 79 and 367, while domain D2 contains 270 amino acids between residues 1171 and 1440 (Figure 4). These domains are separated from each other by a total of 804 residues that contain the triplicated A domains. Domain E1 consists of 46 amino acids between residues 25 and 70, while domain E2 consists of 46 amino acids between residues 1453 and 1498 (Figure

Tuble I: Alignment Scores for Comparisons between Repeated Domains of Human von Willebrand Factor^a

first domain	second domain	alignment score			
(residues)	(residues)	SD units	probability		
A1 (497-716)	A2 (717-909)	10.08	≤10 ⁻²³		
A1 (497-716)	A3 (910-1111)	8.68	≤10 ⁻¹⁶		
A2 (717-909)	A3 (910-1111)	10.59	≤10 ⁻²³		
B1 (1533-1567)	B2 (1577-1602)	3.08	1.04×10^{-3}		
B1 (1533-1567)	B3 (1612-1636)	5.26	7.21×10^{-8}		
B2 (1577-1602)	B3 (1612-1636)	2.95	1.59×10^{-3}		
C1 (1637-1752)	C2 (1781-1899)	10.04	≤10 ⁻²³		
Cx fragment	C1 fragment	5.07	1.99×10^{-7}		
(63-90)	(1661-1692)				
Cx fragment	C2 fragment	6.38	1.00×10^{-10}		
(63-90)	(1812 - 1844)				
D1 (79-367)	D2 (1171-1440)	15.24	≤10 ⁻³⁰		
E1 (25-70)	E2 (1453-1498)	6.00	1.00×10^{-9}		

^aScores are given in standard deviation units and also as estimated probabilities that the proposed alignment could have occurred by chance. Details of the parameters employed for the computer program Align are described in the legend to Figure 4. The number of random runs used to correct for biases introduced by amino acid composition was 100 (Dayhoff et al., 1983).

4). They are separated by 1383 residues that include the triplicated A domains and the duplicated D domains. The region corresponding to the A domains contains only six Cys residues, while the flanking B-E domains are extremely Cys rich, containing a total of 163 Cys. Finally, a short segment (domain Cx fragment, residues 63-90) is very similar to small segments of both domain C1 (residues 1661-1692) and domain C2 (residues 1812-1844). Because this segment overlaps with sequences that we have chosen to include in domains D1 and E1 and because it is very short compared to the other C domains, we have not included it in Figure 4. However, it might represent a remnant of another member of the C-domain family. The segment comparison scores in standard deviation units for the alignments of all proposed repeated domains are shown in Table I.

Homology of von Willebrand Factor to Other Proteins. The amino acid sequence of von Willebrand factor has been compared to all entries in the NBRF Protein Sequence Database. From the ALIGN program (Dayhoff et al., 1983), the A domains of von Willebrand factor appear to be homologous to a portion of complement factor B. For each alignment of an A domain with residues 230-454 of human complement factor B (Mole et al., 1984), 20-24% of the amino acids are identical, and the alignment scores are of range from 5.01 to 5.72 standard deviation units (probability 2.7×10^{-7} to 5.4×10^{-9}). This segment of complement factor B contains the Arg²³⁴-Lys²³⁵ bond that is cleaved during activation by factor D, five residues of the resulting factor Ba peptide, and the aminoterminal 220 residues of the factor Bb polypeptide that precede the serine protease domain. The gene structure of complement factor B indicates that the region encoding this polypeptide contains five intron-exon boundaries at residues 228, 274, 320, 364, and 398 (Campbell et al., 1984). Domains A1 and A2 of von Willebrand factor are encoded together by a single exon; however, domain A3 (residues 910-1111) is interrupted by intron-exon boundaries at amino acid residues 922, 961, 1008, 1056, and 1111 (J. M. Sorace and J. E. Sadler, unpublished results). There is a rough correlation between the location of these intron-exon boundaries in both genes. Twenty-five residues of the corresponding segment of complement component C2a have been reported (Parkes et al., 1983). This sequence also appears to align well with the von Willebrand factor A domains, although the sequence for C2a is too short to evaluate in detail. The biological significance of these

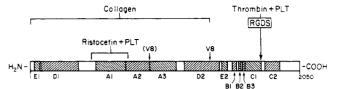


FIGURE 5: Structure-function relationships for human von Willebrand factor. The numbers below the schematic subunit indicate the amino-terminal Ser (1) and the carboxy-terminal Lys (2050). The two principal sites of staphylococcal protease V8 cleavage are indicated by arrows, with the minor site in parentheses. The shaded areas show the extent of internal sequence duplications, or domains, as discussed in the text. The location of a peptide fragment that binds to platelet GPIb is shown by Ristocetin + PLT. The position of the Arg-Gly-Asp-Ser tetrapeptide (RGDS) that may mediate binding to the GPIIb/IIIa complex of activated platelets is shown by Thrombin + PLT. The collagen binding activity of von Willebrand factor is localized to the amino-terminal major V8 protease fragment, as shown.

potential distant relationships is not known.

Aside from these two candidate proteins, the amino acid sequence of von Willebrand factor shows no apparent homology to any protein in the NBRF Protein Sequence Database, and the nucleic acid sequence of the four cDNA isolates is not homologous to any sequence in the Genbank Genomic Sequence Data Bank.

Structure-Function Relationships of von Willebrand Factor. The sequence of the von Willebrand factor precursor surrounding the amino-terminal Ser of the form found in plasma resembles the posttranslational cleavage sites for many other protein precursors. Common features include a Lys-Arg or Arg-Arg dipeptide on the amino-terminal side of the cleavage site. These comparisons suggest that quite distantly related eukaryotic organisms might use similar mechanisms to generate mature secreted protein products. The processing of the von Willebrand factor precursor to its mature form is known to be posttranslational and intracellular and is presumed to occur in the Golgi apparatus (Wagner & Marder, 1983; Lynch et al., 1983). Some proteins, such as proalbumin and many prohormones, are processed intracellularly, too [reviewed in Docherty & Steiner (1982)]. However, despite the similar sequences at the site of protease cleavage, proapolipoprotein A-II is processed extracellularly by a hepatocyte line that processes proalbumin intracellularly (Gordon et al., 1984). Thus, several pathways must exist for the processing of these precursors, even though they share certain structural features. The identity of the required proteases is not known with certainty, although cathepsin B like thiol proteases may participate in both intracellular and extracellular processing reactions (Gordon et al., 1985; Docherty et al., 1982). The kex2 mutants of yeast are deficient in the protease that is required for the analogous processing of the mating factor α precursor (Julius et al., 1984). So far, no homologue for this enzyme has been found in higher eukaryotes.

Some proposed structure—function relationships for mature von Willebrand factor are summarized in Figure 5. Staphylococcal protease V8 cleaves the protein into two major fragments, SPII and SPIII. Fragment SPII is a homodimer of 100-kilodalton peptides containing the carboxy terminus of the subunit, while SPIII is a homodimer of 170-kilodalton amino-terminal peptides. The amino-terminal fragment SPIII retains the ability to bind to both collagen (Fresinaud et al., 1985) and platelet glycoprotein Ib (Girma et al., 1986), and the carboxy-terminal fragment SPII binds to the glycoprotein IIb/IIIa complex of thrombin-activated platelets. The collagen binding activity has not been further localized, but a dimer of 50-kilodalton tryptic fragments of von Willebrand factor has been shown to bind to platelet glycoprotein Ib in the

presence or absence of ristocetin, and this activity is stable to reduction (Fujimura et al., 1986). By protein sequencing, this fragment has been identified as residues 449–729 of the von Willebrand factor subunit. This segment includes all of domain A1, suggesting that at least a part of the binding to glycoprotein Ib is mediated by this small portion of the von Willebrand factor subunit. This fragment was isolated as a dimer from tryptic digests of unreduced von Willebrand factor, and since there are only seven Cys residues in this segment, at least one of them must participate in an intersubunit disulfide bond. Preliminary characterization of the von Willebrand factor gene shows that this functional segment may be encoded by a single exon that species amino acid residues 463–921, including all of domains A1 and A2 (J. M. Sorace and J. E. Sadler, unpublished results).

The tetrapeptide Arg-Gly-Asp-Ser occurs between amino acid residues 1744 and 1747. This sequence also occurs in fibronectin and fibrinogen and appears to mediate the binding of these proteins to the GPIIb/IIIa complex of activated platelets (Pierschbacher & Ruoslahti, 1984a,b). Synthetic peptides containing this sequence inhibit the binding of fibrinogen, fibronectin, and von Willebrand factor to activated platelets (Gartner & Bennett, 1985; Haverstick et al., 1985; Plow et al., 1985). Thus, this small segment of von Willebrand factor may be required for binding to the GPIIb/IIIa complex of activated platelets. It has been placed tentatively at the carboxy-terminal end of domain C1 (Sadler et al., 1985). A factor VIII binding site has not been localized within the linear sequence of von Willebrand factor.

One challenge in the study of proteins with highly repetitive structures, such as fibronectin or von Willebrand factor, is to correlate domains that have been identified by examination of the linear sequence with higher orders of structure such as disulfide bonding patterns, with the organization of the corresponding genomic DNA sequences into introns and exons, and with specific biological functions. Future studies will extend these relationships for von Willebrand factor.

ACKNOWLEDGMENTS

We thank Earl W. Davie for his interest and encouragement, Torben E. Petersen for helpful discussions concerning sequence comparisons, and John R. Bell for his help in synthesizing oligonucleotides.

Registry No. von Willebrand factor, 9001-27-8.

REFERENCES

- Benton, W. D., & Davis, R. W. (1977) Science (Washington, D.C.) 196, 180-182.
- Browning, P. J., Ling, E. H., Zimmerman, T. S., & Lynch, D. C. (1983) Blood 62, 281 (Abstr.).
- Campbell, R. D., Bentley, D. R., & Morley, B. J. (1984) *Philos. Trans. R. Soc. London, B 306*, 367-378.
- Dayhoff, M. O., Barker, W. C., & Hunt, L. T. (1983) Methods Enzymol. 91, 524-545.
- Docherty, K., & Steiner, D. F. (1982) Annu. Rev. Physiol. 44, 625-638.
- Docherty, K., Carroll, R. J., & Steiner, D. F. (1982) Proc. Natl. Acad. Sci. U.S.A. 79, 4613-4617.
- Fay, P. J., Kawai, Y., Wagner, D. D., Ginsburg, D., Bonthron,
 D., Ohlsson-Wilhelm, B. M., Chavin, S. I., Abraham, G.
 N., Handin, R. I., Orkin, S. H., Montgomery, R. R., &
 Marder, V. J. (1985) Blood 66(Suppl. 1), 333a.
- Fujimura, Y., Titani, K., Holland, L. Z., Russell, S. R., Roberts, J. R., Elder, J. H., Ruggeri, Z. M., & Zimmerman, T. S. (1986) J. Biol. Chem. 261, 381-385.

- Gartner, T. K., & Bennett, J. S. (1985) J. Biol. Chem. 260, 11891-11894.
- Ginsburg, D., Handin, R. I., Bonthron, D. T., Donlon, T. A., Bruns, G. A. P., Latt, S. A., & Orkin, S. H. (1985) *Science* (*Washington*, *D.C.*) 228, 1401-1406.
- Girma, J.-P., Pietu, G., Chopek, M. W., Edgington, T. S., & Meyer, D. (1984) Circulation 70(Suppl. II), Abstr. 836.
- Girma, J.-P., Chopek, M. W., Titani, K., & Davie, E. W. (1986) *Biochemistry* (second paper of four in this issue).
- Gordon, J. I., Sims, H. F., Edelstein, C., Scanu, A. M., & Strauss, A. W. (1984) J. Biol. Chem. 259, 15556-15563.
- Gordon, J. I., Sims, H. F., Edelstein, C., Scanu, A. M., & Strauss, A. W. (1985) J. Biol. Chem. 260, 14824-14831.
- Haverstick, D. M., Cowan, J. F., Yamada, K. M., & Santoro, S. A. (1985) *Blood 66*, 946-952.
- Hoyer, L. W. (1981) Blood 58, 1-13.
- Jaffe, E. A., & Hoyer, L. W. (1974) Proc. Natl. Acad. Sci. U.S.A. 71, 1906-1909.
- Jaffe, E. A., Hoyer, L. W., & Nachman, R. L. (1973) J. Clin. Invest. 52, 2757-2764.
- Jenkins, C. S. P., Philips, D. R., Clemetson, K. J., Meyer, D., Larrieu, M.-J., & Luscher, E. F. (1976) J. Clin. Invest. 57, 112-124.
- Julius, D., Brake, A., Blair, L., Kunisawa, R., & Thorner, J. (1984) Cell (Cambridge, Mass.) 37, 1074-1089.
- Kornblihtt, A. R., Umezawa, K., Vibe-Pedersen, K., & Baralle, F. E. (1985) *EMBO J. 4*, 1755-1759.
- Lathe, R. (1985) J. Mol. Biol. 183, 1-12.
- Lynch, D. C., Williams, R., Zimmerman, T. S., Kirby, E. P., & Livingston, D. M. (1983) Proc. Natl. Acad. Sci. U.S.A. 80, 2738-2742.
- Lynch, D. C., Zimmerman, T. S., Collins, C. J., Brown, M., Morin, M. J., Ling, E. H., & Livingston, D. M. (1985) Cell (Cambridge, Mass.) 41, 49-56.
- Maniatis, T., Jeffrey, A., & Kleid, D. G. (1975) Proc. Natl. Acad. Sci. U.S.A. 72, 1184–1188.
- Mole, J. E., Anderson, J. K., Davison, E. A., & Woods, D.E. (1984) J. Biol. Chem. 259, 3407-3412.
- Montgomery, R. R., & Zimmerman, T. S. (1978) J. Clin. Invest. 61, 1498-1507.
- Morton, L. F., Griffin, B., Pepper, D. S., & Barnes, M. J. (1983) *Thromb. Res.* 32, 545-556.
- Nachman, R. L., Levine, R., & Jaffe, E. A. (1977) J. Clin. Invest. 60, 914-921.
- Parkes, C., Gagnon, J., & Kerr, M. A. (1983) *Biochem. J.* 213, 201-209.
- Pierschbacher, M. D., & Ruoslahti, E. (1984a) Nature (London) 309, 30-33.
- Pierschbacher, M. D., & Ruoslahti, E. (1984b) *Proc. Natl. Acad. Sci. U.S.A.* 81, 5985-5988.
- Pietu, G., Cherel, G., Marguerie, G., & Meyer, D. (1984)

 Nature (London) 308, 648-649.
- Plow, E. F., Pierschbacher, M. D., Ruoslahti, E., Marguerie, G. A., & Ginsburg, M. H. (1985) Proc. Natl. Acad. Sci. U.S.A. 82, 8057-8061.
- Sadler, J. E., Shelton-Inloes, B. B., Sorace, J. M., Harlan, J. M., Titani, K., & Davie, E. W. (1985) *Proc. Natl. Acad. Sci. U.S.A.* 82, 6394-6398.
- Santoro, S. A. (1981) Thromb. Res. 21, 689-693.
- Santoro, S. A., & Cowan, J. F. (1982) Collagen Relat. Res. 2, 31-43.
- 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.
 Thomas, P. S. (1983) Methods Enzymol. 100, 255-266.

Titani, K., Kumar, S., Takio, K., Ericsson, L. H., Wade, R. D., Ashida, K., Walsh, K. A., Chopek, M. W., Sadler, J. E., & Fujikawa, K. (1986) *Biochemistry* (fourth paper of four in this issue).

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.

Verweij, C. L., de Vries, C. J. M., Distel, B., van Zonneveld, A.-J., van Kessel, A. G., van Mourik, J. A., & Pannekoek,

H. (1985) Nucleic Acids Res. 13, 4699-4717.

Wagner, D. D., & Marder, V. J. (1983) J. Biol. Chem. 258, 2065-2067.

Wagner, D. D., & Marder, V. J. (1984) J. Cell Biol. 99, 2123-2130.

Wagner, D. D., Urban-Pickering, M., & Marder, V. J. (1984) *Proc. Natl. Acad. Sci. U.S.A.* 81, 471-475.

Wilbur, W. J., & Lipman, D. J. (1983) Proc. Natl. Acad. Sci. U.S.A. 80, 726-730.

Amino Acid Sequence of Human von Willebrand Factor[†]

Koiti Titani,*,† Santosh Kumar,† Koji Takio,† Lowell H. Ericsson,† Roger D. Wade,† Katsuro Ashida,† Kenneth A. Walsh,† Michael W. Chopek,†,§ J. Evan Sadler, and Kazuo Fujikawa†

Department of Biochemistry, University of Washington, Seattle, Washington 98195, and Departments of Medicine and Biochemistry, Washington University School of Medicine, St. Louis, Missouri 63110

Received January 14, 1986; Revised Manuscript Received February 25, 1986

ABSTRACT: The complete amino acid sequence of human von Willebrand factor (vWF) is presented. Most of the sequence was determined by analysis of the S-carboxymethylated protein. Some overlaps not provided by the protein sequence analysis were obtained from the sequence predicted by the nucleotide sequence of a cDNA clone [Sadler, J. E., Shelton-Inloes, B. B., Sorace, J., Harlan, M., Titani, K., & Davie, E. W. (1985) Proc. Natl. Acad. Sci. U.S.A. 82, 6391-6398]. The protein is composed of 2050 amino acid residues containing 12 Asn-linked and 10 Thr/Ser-linked oligosaccharide chains. One of the carbohydrate chains is linked to an Asn residue in the sequence Asn-Ser-Cys rather than the usual Asn-X-Ser/Thr sequence. The sequence of von Willebrand factor includes several regions bearing evidence of internal gene duplication of ancestral sequences. The protein also contains the tetrapeptide sequence Arg-Gly-Asp-Ser (at residues 1744-1747), which may be a cell attachment site, as in fibronectin. The amino- and carboxyl-terminal regions of the molecule contain clusters of half-cystinyl residues. The sequence is unique except for some homology to human complement factor B.

Human von Willebrand factor (vWF)¹ is a plasma glycoprotein (Legaz et al., 1973; Shapiro et al., 1973; Olson et al., 1977) that is involved in platelet adhesion to the subendothelium, leading to platelet plug formation during vascular injury (Jorgensen & Borchgrevin, 1964; Havig & Stormoken, 1974). The prolonged bleeding time of individuals having low levels of vWF or modified vWF is due to poor platelet plug formation (Ruggeri et al., 1982; Hoyer, 1982; Kinoshita et al., 1984).

vWF is synthesized in endothelial cells (Jaffe et al., 1973; Jaffe & Hoyer, 1974) and megakaryocytes (Nachman et al., 1977) in a large precursor form and secreting into plasma after several processing events, including glycosylation, sulfation, disulfide formation, and proteolytic cleavages (Wagner & Marder, 1983, 1984; Browning et al., 1983; Lynch et al., 1983; Ling et al., 1984). It circulates in plasma as a series of high molecular weight multimers ranging in size from 1×10^6 to 12×10^6 daltons (Counts et al., 1978; Perret et al., 1979;

Ruggeri & Zimmerman, 1980; Hoyer & Shainoff, 1980; Meyer et al., 1980). Electron micrographs suggest that extended protomers of 100–120 nm in length assemble into the multimeric structures that circulate in plasma (Slayter et al., 1985; Fowler et al., 1985).

We have established a large-scale purification procedure for human vWF from a commercial factor VIII concentrate and presented preliminary evidence that it is composed of identical subunits of approximate M_r 270 000. This conclusion was based on the observation of a single amino acid sequence at the amino terminus as well as at the carboxyl terminus. This was confirmed by the agreement between the number of unique cyanogen bromide fragments and that predicted from the methionine content of the protein (Chopek et al., 1986). We have also shown that limited proteolysis of native vWF by Staphylococcus aureus V8 protease produces two major fragments that can be separated without cleaving the disulfide bonds. One of these (fragment III) is a 170K-dalton segment from the amino terminus that retains binding activity to ristocetin-treated platelets (Girma et al., 1986). The other (fragment II) is a 100K-dalton fragment from the carboxylterminal end of the protein. It binds platelets activated by either ADP or thrombin (Girma et al., 1984). A third (minor) fragment (fragment I) is a 50K-dalton subdigestion product of fragment III.

[†]This work was supported by Research Grants HL29595 (to K. Titani) and HL16919 (to E.W.D.) from the National Institutes of Health. K. Titani is a visiting professor of the Fujita-Gakuen Health University, Nagoya, Japan. K. Takio is a senior associate of the Howard Hughes Medical Institute Laboratory in Seattle. J.E.S. is an Associate Investigator of the Howard Hughes Medical Institute in Washington University, St. Louis, MO.

[‡]University of Washington.

[§] Present address: St. Paul Regional Red Cross, St. Paul, MN 55107, or Department of Laboratory Medicine and Pathology, University of Minnesota, Minneapolis, MN 55455.

Washington University School of Medicine.

¹ Abbreviations: vWF, von Willebrand factor; HPLC, high-performance liquid chromatography; CM, carboxymethyl; RP, reversed phase; Tris-HCl, tris(hydroxymethyl)aminomethane hydrochloride.