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 270000. 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 carboxyl-terminal 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.

0006-2960/86/0425-3171$01.50/0

Sadler et al. (1985) have recently isolated two cDNA clones coding for about 80% of human vWF. In these studies, a cDNA library was prepared in λgt11 bacteriophage with poly(A) RNA isolated from primary cultures of human endothelial cells. The library was screened with affinity-purified antibody, and clones were isolated that contained cDNA inserts that correlated with our amino acid sequence data (Chopek et al., 1986; Girma et al., 1986). One clone (λHvWF1) contained an insert of 404 nucleotides that corresponded to amino acid residues 1–110 in the mature protein and defined a 24-residue segment of a precursor leader sequence. The second cDNA clone (λHvWF3) contained a 4.9-kilobase insert that corresponded to the carboxyl-terminal 1525 amino acid residues of the protein, followed by the stop codon TGA, 134 nucleotides of 3′-noncoding sequence, and a poly(A) tail of 150 nucleotides. These two cDNA inserts did not overlap, and the present study provides the intervening segment. Three other reports have recently described the isolation of cDNA clones coding for segments of human vWF (Lynch et al., 1985; Ginsburg et al., 1985; Verweij et al., 1985). In each of these cases, unpublished amino acid sequence data from our laboratory were used to establish that these clones encoded vWF.

In this paper, amino acid sequence data obtained by the direct analysis of human vWF is presented. We describe the amino acid sequence of 2050 residues, derived by a combination of protein and cDNA sequencing (Sadler et al., 1985). It is now possible to explore the relationship of unique structural features of vWF to its functional substructural domains.

Materials and Methods

Human vWF was purified from a commercial factor VIII concentrate (a generous gift of Dr. H. Kingdon, Hyland Therapeutics, Division of Travenol Laboratories) by a slight modification of the method previously described (Chopek et al., 1986). The factor VIII concentrate (750 mL, 23 mg of protein/mL) was subjected to gel filtration on a column of Sepharose CL-4B equilibrated with 20 mM imidazole hydrochloride buffer, pH 6.8. The break-through peak was passed through a gelatin–Sepharose column and then an agmatine–Sepharose column as previously described (Chopek et al., 1986). vWF was precipitated by adjusting the final concentration to 40% saturation of $(NH_4)_2SO_4$. The precipitate was stored at −20 °C. This modified procedure yields ca. 300 mg of the protein from 750 mL of the factor VIII concentrate in two days.

Fragments I–III were prepared by limited proteolysis of the native protein with *Staphylococcus aureus* V8 protease (Miles), as previously described by Girma et al. (1986), but with the following modifications. The purified vWF was dialyzed against 0.1 M NH_4HCO_3, pH 8.0, containing 0.01% sodium azide, at 4 °C for 2 days, and incubated with the protease at a molar ratio of 1:100 at 25 °C for 48 h. The digest was applied to a Mono Q column (0.5 × 5 cm, Pharmacia) equilibrated with 50 mM Tris-HCl, pH 7.4, and eluted by a linear gradient of NaCl as shown in Figure 1. Three fractions (fragments I–III) were pooled and stored at 5 °C. For the sequence analysis, the fragments were precipitated by the addition of trichloroacetic acid before reduction and S-carboxymethylation. For separate binding studies, the fragments were not precipitated but dialyzed against physiological buffer before use.

TPCK-trypsin and α-chymotrypsin were purchased from Worthington, and lima bean trypsin inhibitor (LBTI) was from Millipore. *Achromobacter* protease, which specifically cleaves

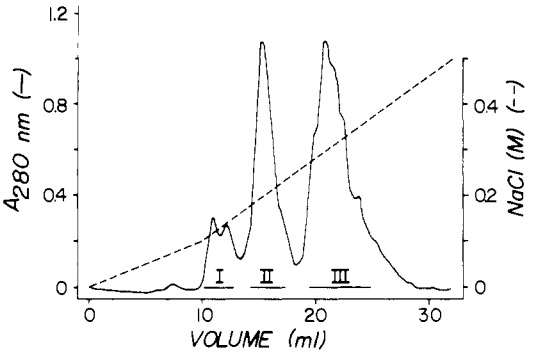


FIGURE 1: Separation of *S. aureus* V8 fragments I–III of human vWF on a Mono Q column. Fifteen milligrams of *S. aureus* V8 digest of native vWF in 0.1 M ammonium bicarbonate (40 h) was applied to a column (0.5 × 5 cm) of Mono Q, equilibrated with 20 mM Tris-HCl, pH 7.4, and eluted at room temperature with a linear gradient from 0 to 0.5 M NaCl in the same buffer. Fragments I–III were detected by absorbance at 280 nm and pooled as indicated by horizontal bars.

lysyl bonds (Masaki et al., 1981), was a generous gift of Dr. T. Masaki of the Department of Agricultural Chemistry, Ibaraki University, Japan. [^{14}C]Iodoacetic acid, [^{3}H]iodoacetic acid, and [^{14}C]methyl iodide were products of New England Nuclear.

Amino acid compositions of the large fragments were determined with a Dionex D-500 instrument. Smaller peptides, particularly in the digest by the *Achromobacter* protease, were analyzed by the Waters picotag method of Bidlingmeyer et al. (1984). Edman degradations were performed in a Beckman 890C spinning-cup sequencer with >2 nmol of peptide (Takio et al., 1982) or in an Applied Biosystems sequencer with smaller amounts (Hewick et al., 1981). Phenylthiohydantoins were identified by the method of Ericsson et al. (1977) on a Zorbax ODS column and confirmed on a (cyanopropyl)silica column by the method of Hunkapiller and Hood (1983) or more recently on the Zorbax PTH column of Glajch et al. (1985).

Methods employed for reduction, carboxymethylation, [^{14}C]methylation of methionine, citraconylation, and digestion with enzymes or cyanogen bromide followed procedures previously described (Titani et al., 1984). Complex mixtures of peptides were usually separated first by gel filtration on Sephadex G-75 or by HPLC on TSK G3000 SW columns. The larger peptides were then purified by RP-HPLC on Ultrapore RPSC (C3), SynChropak RP-8 (C8), or Cosmosil 5C18P (C18) columns and the smaller peptides on a SynChropak RP-P (C18) column. In each case, an acetonitrile gradient was used in dilute aqueous trifluoroacetic acid.

The hydropathy index and probable antigenic sites were examined by the methods of Kyte and Doolittle (1982) and Hopp and Woods (1981), respectively, with a seven-residue moving average.

Results

General Strategy of Sequence Analysis. A complete set of 42 fragments was generated from the whole CM protein by cleavage with cyanogen bromide at methionyl bonds. These fragments were grouped by their placement within two large segments (fragments II and III), obtained by limited proteolysis with the *S. aureus* V8 protease. The sequences of the methionyl fragments were aligned by using peptides isolated from two digests, one by cleavage at lysyl bonds with *Achromobacter* protease I and the other by cleavage at arginyl bonds with trypsin after blocking ε-amino groups of the CM protein by citraconylation. This general strategy is illustrated

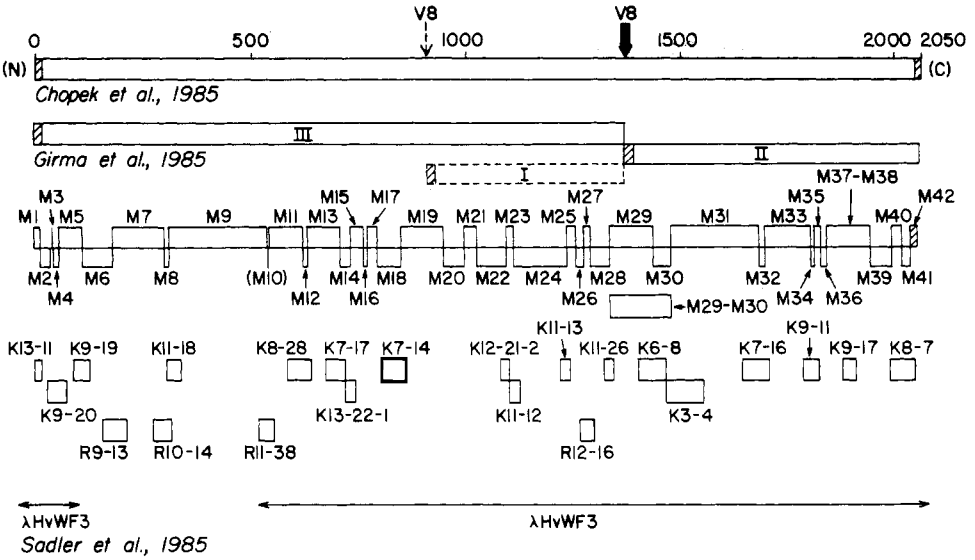


FIGURE 2: Relative orientation of the major fragments of human vWF. Fragments I–III of Girma et al. (1986) refer to the products of limited proteolysis by *S. aureus* V8 protease. Peptides with prefixes M, K, and R are products of cleavage at methionyl, lysyl, and arginyl bonds, respectively. Hashed portions of bars indicate sequences published previously from our laboratories. Each peptide is designated by a bar of a length consistent with the 2050-residue whole polypeptide chain at the top. Details of the structural analyses of these and other peptides are illustrated in Figure 3. The alignment of the cDNA sequence segments of Sadler et al. (1985) is indicated at the bottom.

in Figure 2, which displays the relative orientation of the peptides and the location of the four short sequences previously described at the termini of fragments II and III (Chopek et al., 1986; Girma et al. 1986). More complete details of the proof of sequence are summarized in Figure 3.

Cleavage with Cyanogen Bromide. Figure 4 illustrates the primary step in the separation of the peptides generated by cleavage with cyanogen bromide of the whole ^{3}H-labeled CM protein (25 mg) and of the *S. aureus* V8 fragments III (10–15 mg) and II (5–10 mg) (after reduction and S-carboxymethylation). Digestion of the whole CM protein with CNBr was repeated 3 more times (25 mg each time) to provide adequate amounts of fragments for the sequence analysis. Further purification of the peptides in each pool was carried out by RP-HPLC. Initially for large peptides, C3 or C8 columns (e.g., Ultrapore RPSC or SynChropak RP-8) were used. More recently, 5-μm C18 protein columns (e.g., Cosmosil 5C18P) were employed for this purpose. A C18 column (e.g., SynChropack RP-P) was used throughout for small peptides as illustrated in Figure 5.

It was difficult to separate the complete set of cyanogen bromide fragments from a digest of the complexity arising from a protein of this size. For example, in early experiments the two largest fragments, M9 and M31, eluted together in pool 1 (Figure 4A), and their subsequent separation was in very poor yield (<5%) on an Altex Ultrapore RPSC (C3) column. This problem was circumvented by isolating these two fragments from separate cyanogen bromide digests of proteolytic fragments III and II. Thus, pool 1 in Figure 4B contained only fragment M9, and pool 1 in Figure 4C contained only fragment M31. These fragments were then desalted on a small Sephadex G-25 column in 9% formic acid. Similarly, peptides M24 and M33 in pool 3 (Figure 4A) were not sufficiently resolved on any of the RP-HPLC columns tested but were readily isolated from pool 2 in Figure 4B and pool 2 in Figure 4C, respectively, by RP-HPLC.

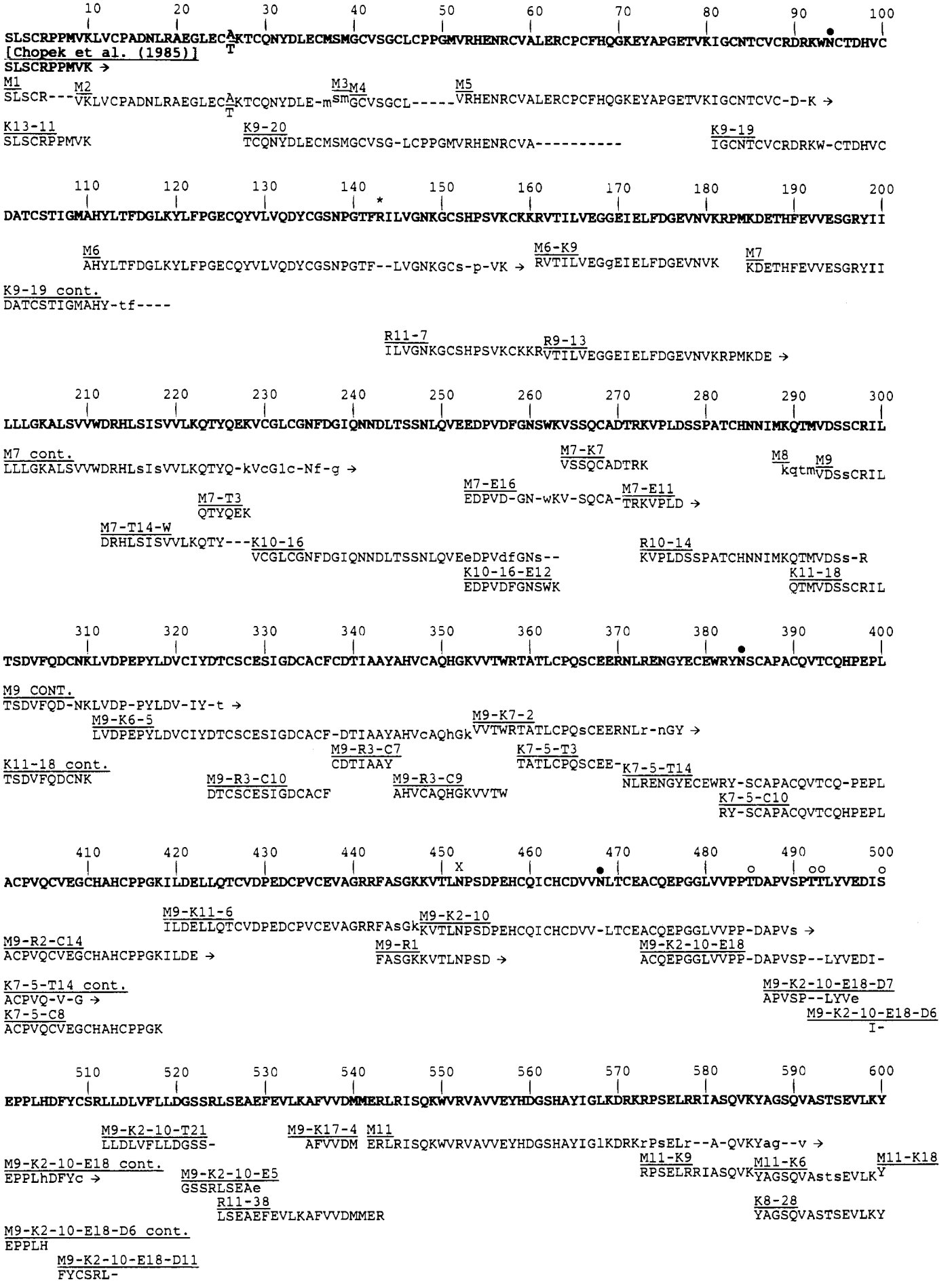
A complete set of the fragments except for M10 (free homoserine) was isolated in this manner from the three cyanogen bromide digests. Incomplete cleavage at a Met–Ser bond (residues 1438–1439) yielded an overlapping fragment, M29-M30. A Met–Thr bond at residues 1896–1897 was quite resistant to cyanogen bromide, yielding primarily the overlapping fragment M37-M38, although small amounts of M37 and M38 were isolated from a digest of fragment II. Free homoserine should have been released from residue 541 (Figure 3), but it was not found during the fractionation of these digests. It is nonetheless referred to as peptide M10.

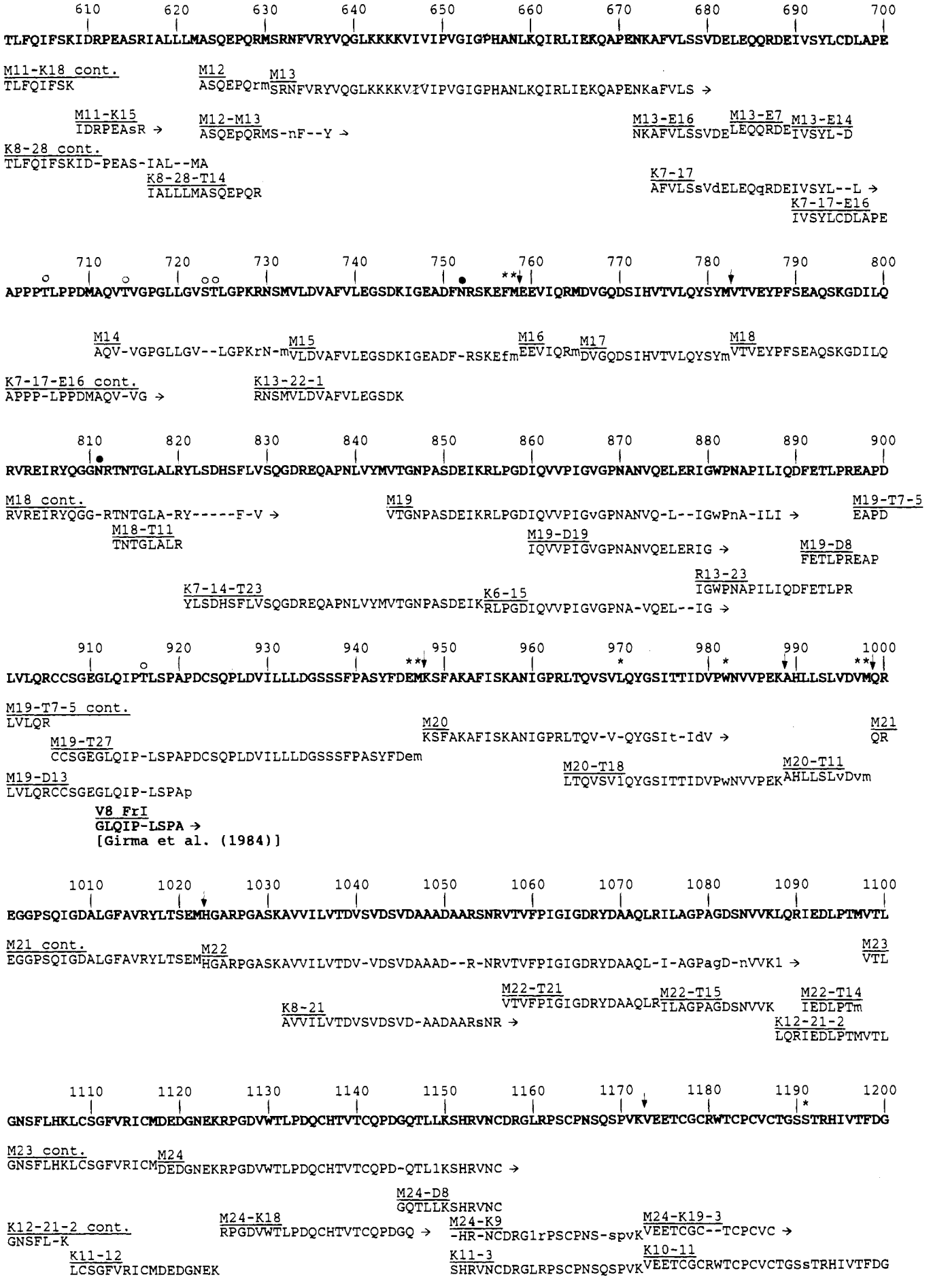
Limited proteolysis of the native protein by *S. aureus* V8 protease (producing fragments III and II; cf. Figure 1) occurred mainly in the middle of the sequence corresponding to M29 (between residues 1365 and 1366). Subsequent digestion of the separated fragments III and II with cyanogen bromide generated two additional major fragments M29(N) and M29(C), as well as an overlap fragment, M29(C)-M30. These were not present in the digest of the whole CM-protein.

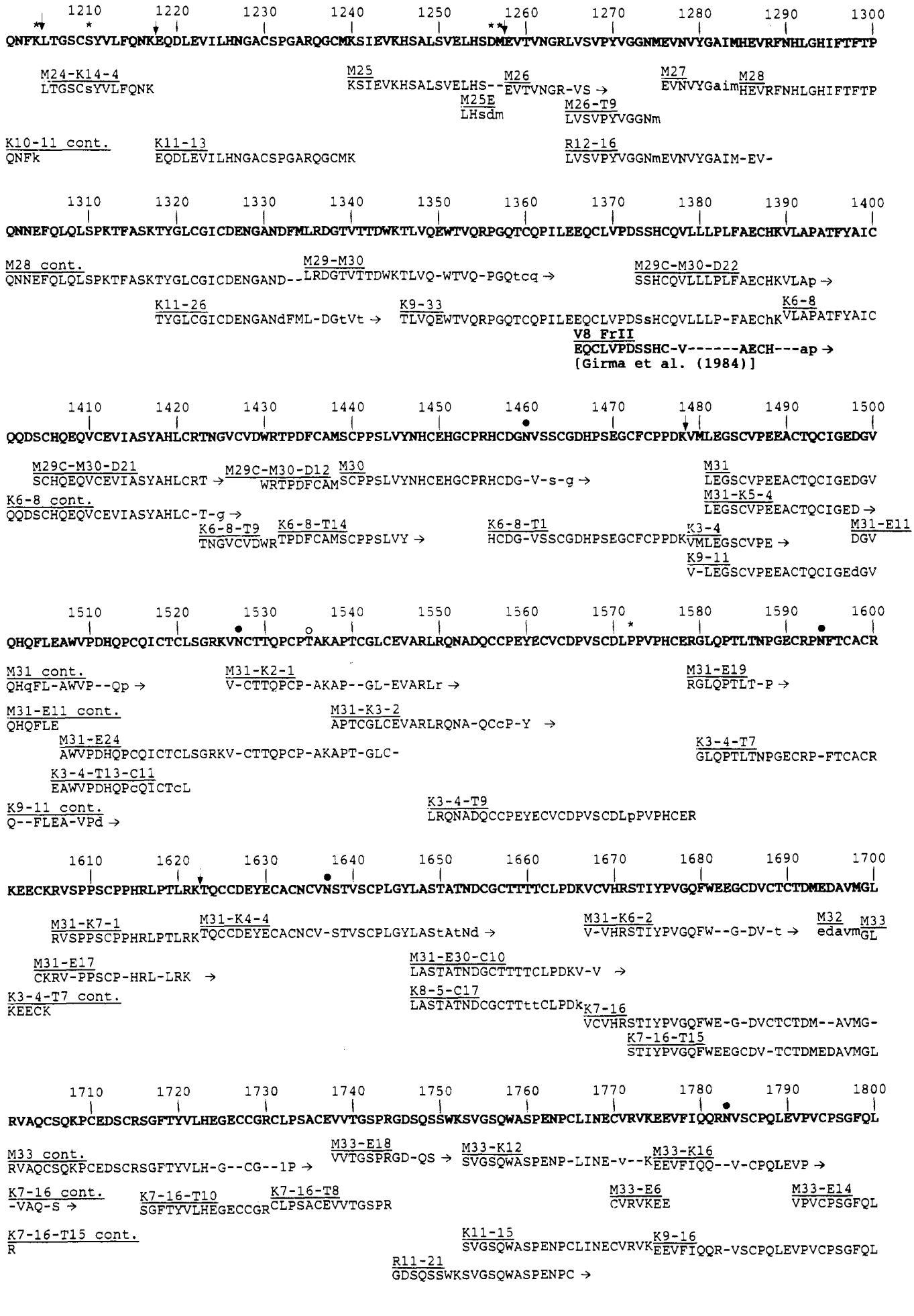
The amino acid compositions calculated from the sequences of the unique fragments M1 through M42 (excluding overlapping fragment M29-M30 and fragments M29(N), M29(C), and M29(C)-M30 generated by *S. aureus* V8 protease) was reasonably consistent with the amino acid composition in hydrolysates of the whole protein, calculated with a M_r of 270K-daltons and a carbohydrate content of 15% (Table I).

Of the cyanogen bromide fragments isolated and analyzed, M1 (eight residues) had an amino-terminal sequence identical with that of the whole protein, Ser-Leu-Ser-Cys-Arg, indicating that this fragment was derived from the amino-terminal portion of the protein. Moreover, only M42 (11 residues) lacked homoserine, indicating that this fragment originated from the carboxyl terminus of the protein (Chopek et al., 1986). Of the fragments, only 11 (M12, M15, M16, M17, M21, M23, M27, M35, M36, M40, and M42) were sequenced to their carboxyl termini. The others required subdigestion by trypsin, chymotrypsin, *S. aureus* V8 protease, *Achromobacter* protease, and/or dilute acid to extend the sequences as shown in Figure 3. Fragments of five residues or less (M3, M8, M10, M32, and M34) were not analyzed directly, but their sequences were later identified by analyses of the peptides generated by digestion of the whole protein with *Achromobacter* protease or trypsin.

In all, sequenator analyses of the cyanogen bromide fragments and subpeptides derived therefrom placed about 80% of the 2050 amino acid residues of the whole protein in


SLSCRPPMVKLVCPADNLRAEGLECA/TKTCQNYDLECMSMGCVSGCLCPPGMVRHENRCVALERCPCFHQGKEYAPGETVKIGCNTCVCRDRKWNCTDHVC
[Chopek et al. (1985)]
SLSCRPPMVK →
M1 SLSCR---
M2 VKLVCPADNLRAEGLECA/TKTCQNYDLE-m
M3 sm M4 GCVSGCL-----
M5 VRHENRCVALERCPCFHQGKEYAPGETVKIGCNTCVC-D-K →
K13-11 SLSCRPPMVK
K9-20 TCQNYDLECMSMGCVSG-LCPPGMVRHENRCVA----------
K9-19 IGCNTCVCRDRKW-CTDHVC
DATCSTIGMAHYLTFDGLKYLFPGECQYVLVQDYCGSNPGTFRILVGNKGCSHPSVKCKKRVTILVEGGEIELFDGEVNVKRPMKDETHFEVVESGRYII
M6 AHYLTFDGLKYLFPGECQYVLVQDYCGSNPGTF--LVGNKGCs-p-VK →
M6-K9 RVTILVEGgEIELFDGEVNVK
M7 KDETHFEVVESGRYII
K9-19 cont. DATCSTIGMAHY-tf----
R11-7 ILVGNKGCSHPSVKCKKR
R9-13 VTILVEGGEIELFDGEVNVKRPMKDE →
LLLGKALSVVWDRHLSISVVLKQTYQEKVCGLCGNFDGIQNNDLTSSNLQVEEDPVDFGNSWKVSSQCADTRKVPLDSSPATCHNNIMKQTMVDSSCRIL
M7 cont. LLLGKALSVVWDRHLsIsVVLKQTYQ-kVcGlc-Nf-g →
M7-K7 VSSQCADTRK
M8 kqtm
M9 VDSscRIL
M7-E16 EDPVD-GN-wKV-SQCA-
M7-E11 TRKVPLD →
M7-T3 QTYQEK
M7-T14-W DRHLSISVVLKQTY---
K10-16 VCGLCGNFDGIQNNDLTSSNLQVEeDPVdfGNs--
K10-16-E12 EDPVDFGNSWK
R10-14 KVPLDSSPATCHNNIMKQTMVDSs-R
K11-18 QTMVDSSCRIL
TSDVFQDCNKLVDPEPYLDVCIYDTCSCESIGDCACFCDTIAAYAHVCAQHGKVVTWRTATLCPQSCEERNLRENGYECEWRYNSCAPACQVTCQHPEPL
M9 CONT. TSDVFQD-NKLVDP-PYLDV-IY-t →
M9-K6-5 LVDPEPYLDVCIYDTCSCESIGDCACF-DTIAAYAHVcAQhGk
M9-K7-2 VVTWRTATLCPQsCEERNLr-nGY →
M9-R3-C7 CDTIAAY
K7-5-T3 TATLCPQSCEE-
K11-18 cont. TSDVFQDCNK
M9-R3-C10 DTCSCESIGDCACF
M9-R3-C9 AHVCAQHGKVVTW
K7-5-T14 NLRENGYECEWRY-SCAPACQVTCQ-PEPL
K7-5-C10 RY-SCAPACQVTCQHPEPL
ACPVQCVEGCHAHCPPGKILDELLQTCVDPEDCPVCEVAGRRFASGKKVTLNPSDPEHCQICHCDVVNLTCEACQEPGGLVVPPTDAPVSPTTLYVEDIS
M9-K11-6 ILDELLQTCVDPEDCPVCEVAGRRFAsGk
M9-K2-10 KVTLNPSDPEHCQICHCDVV-LTCEACQEPGGLVVPP-DAPVs →
M9-R2-C14 ACPVQCVEGCHAHCPPGKILDE →
M9-R1 FASGKKVTLNPSD →
M9-K2-10-E18 ACQEPGGLVVPP-DAPVSP--LYVEDI-
K7-5-T14 cont. ACPVQ-V-G →
K7-5-C8 ACPVQCVEGCHAHCPPGK
M9-K2-10-E18-D7 APVSP--LYVe
M9-K2-10-E18-D6 I-
EPPLHDFYCSRLLDLVFLLDGSSRLSEAEFEVLKAFVVDMMERLRISQKWVRVAVVEYHDGSHAYIGLKDRKRPSELRRIASQVKYAGSQVASTSEVLKY
M9-K2-10-T21 LLDLVFLLDGSS-
M9-K17-4 AFVVDM
M11 ERLRISQKWVRVAVVEYHDGSHAYIGlKDRKrPsELr--A-QVKYag--v →
M9-K2-10-E18 cont. EPPLhDFYc →
M9-K2-10-E5 GSSRLSEAe
M11-K9 RPSELRRIASQVK
M11-K6 YAGSQVAstsEVLK
M11-K18 Y
R11-38 LSEAEFEVLKAFVVDMMER
K8-28 YAGSQVASTSEVLKY
M9-K2-10-E18-D6 cont. EPPLH
M9-K2-10-E18-D11 FYCSRL-


610 620 630 640 650 660 670 680 690 700
TLFQIFSKIDRPEASRIALLLMASQEPQRMSRNFVRYVQGLKKKKVIVIPVGIGPHANLKQIRLIEKQAPENKAFVLSSVDELEQQRDEIVSYLCDLAPE
M11-K18 cont.
TLFQIFSK
M12
ASQEPQrm
M13
SRNFVRYVQGLKKKKVIVIPVGIGPHANLKQIRLIEKQAPENKaFVLS →
M11-K15
IDRPEAsR →
M12-M13
ASQEpQRMS-nF--Y →
M13-E16
NKAFVLSSVDE
M13-E7
LEQQRDE
M13-E14
IVSYL-D
K8-28 cont.
TLFQIFSKID-PEAS-IAL--MA
K8-28-T14
IALLLMASQEPQR
K7-17
AFVLSsVdELEQqRDEIVSYL--L →
K7-17-E16
IVSYLCDLAPE
710 720 730 740 750 760 770 780 790 800
APPPTLPPDMAQVTVGPGLLGVSTLGPKRNSMVLDVAFVLEGSDKIGEADFNRSKEFMEEVIQRMDVGQDSIHVTVLQYSYMVTVEYPFSEAQSKGDILQ
M14
AQV-VGPGLLGV--LGPKrN-m
M15
VLDVAFVLEGSDKIGEADF-RSKEfm
M16
EEVIQRm
M17
DVGQDSIHVTVLQYSYm
M18
VTVEYPFSEAQSKGDILQ
K7-17-E16 cont.
APPP-LPPDMAQV-VG →
K13-22-1
RNSMVLDVAFVLEGSDK
810 820 830 840 850 860 870 880 890 900
RVREIRYQGGNRTNTGLALRYLSDHSFLVSQGDREQAPNLVYMVTGNPASDEIKRLPGDIQVVPIGVGPNANVQELERIGWPNAPILIQDFETLPREAPD
M18 cont.
RVREIRYQGG-RTNTGLA-RY-----F-V →
M19
VTGNPASDEIKRLPGDIQVVPIGvGPNANVQ-L--IGwPnA-ILI →
M19-T7-5
EAPD
M18-T11
TNTGLALR
M19-D19
IQVVPIGVGPNANVQELERIG →
M19-D8
FETLPREAP
R13-23
IGWPNAPILIQDFETLPR
K7-14-T23
YLSDHSFLVSQGDREQAPNLVYMVTGNPASDEIK
K6-15
RLPGDIQVVPIGVGPNA-VQEL--IG →
910 920 930 940 950 960 970 980 990 1000
LVLQRCCSGEGLQIPTLSPAPDCSQPLDVILLLDGSSSFPASYFDEMKSFAKAFISKANIGPRLTQVSVLQYGSITTIDVPWNVVPEKAHLLSLVDVMQR
M19-T7-5 cont.
LVLQR
M20
KSFAKAFISKANIGPRLTQV-V-QYGSIt-IdV →
M21
QR
M19-T27
CCSGEGLQIP-LSPAPDCSQPLDVILLLDGSSSFPASYFDem
M20-T18
LTQVSVIQYGSITTIDVPwNVVPEK
M20-T11
AHLLSLvDvm
M19-D13
LVLQRCCSGEGLQIP-LSPAp
V8 FrI
GLQIP-LSPA →
[Girma et al. (1984)]
1010 1020 1030 1040 1050 1060 1070 1080 1090 1100
EGGPSQIGDALGFAVRYLTSEMHGARPGASKAVVILVTDVSVDSVDAAADAARSNRVTVFPIGIGDRYDAAQLRILAGPAGDSNVVKLQRIEDLPTMVTL
M21 cont.
EGGPSQIGDALGFAVRYLTSEM
M22
HGARPGASKAVVILVTDV-VDSVDAAAD--R-NRVTVFPIGIGDRYDAAQL-I-AGPagD-nVVK1 →
M23
VTL
M22-T21
VTVFPIGIGDRYDAAQLR
M22-T15
ILAGPAGDSNVVK
M22-T14
IEDLPTm
K8-21
AVVILVTDVSVDSVD-AADAARsNR →
K12-21-2
LQRIEDLPTMVTL
1110 1120 1130 1140 1150 1160 1170 1180 1190 1200
GNSFLHKLCSGFVRICMDEDGNEKRPGDVWTLPDQCHTVTCQPDGQTLLKSHRVNCDRGLRPSCPNSQSPVKVEETCGCRWTCPCVCTGSSTRHIVTFDG
M23 cont.
GNSFLHKLCSGFVRICM
M24
DEDGNEKRPGDVWTLPDQCHTVTCQPD-QTL1KSHRVNC →
M24-D8
GQTLLKSHRVNC
K12-21-2 cont.
GNSFL-K
M24-K18
RPGDVWTLPDQCHTVTCQPDGQ →
M24-K9
-HR-NCDRG1rPSCPNS-spvK
M24-K19-3
VEETCGC--TCPCVC →
K11-12
LCSGFVRICMDEDGNEK
K11-3
SHRVNCDRGLRPSCPNSQSPVK
K10-11
VEETCGCRWTCPCVCTGSsTRHIVTFDG


1210 1220 1230 1240 1250 1260 1270 1280 1290 1300
QNFKLTGSCSYVLFQNKEQDLEVILHNGACSPGARQGCMKSIEVKHSALSVELHSDMEVTVNGRLVSVPYVGGNMEVNVYGAIMHEVRFNHLGHIFTFTP
M24-K14-4 LTGSCsYVLFQNK
M25 KSIEVKHSALSVELHS--
M26 EVTVNGR-VS →
M27 EVNVYGaim
M28 HEVRFNHLGHIFTFTP
M25E LHsdm
M26-T9 LVSVPYVGGNm
K10-11 cont. QNFk
K11-13 EQDLEVILHNGACSPGARQGCMK
R12-16 LVSVPYVGGNmEVNVYGAIM-EV-
1310 1320 1330 1340 1350 1360 1370 1380 1390 1400
QNNEFQLQLSPKTFASKTYGLCGICDENGANDFMLRDGTVTTDWKTLVQEWTVQRPGQTCQPILEEQCLVPDSSHCQVLLLPLFAECHKVLAPATFYAIC
M28 cont. QNNEFQLQLSPKTFASKTYGLCGICDENGAND--
M29-M30 LRDGTVTTDWKTLVQ-WTVQ-PGQtcq →
M29C-M30-D22 SSHCQVLLLPLFAECHKVLAp →
K11-26 TYGLCGICDENGANdFML-DGtVt →
K9-33 TLVQEWTVQRPGQTCQPILEEQCLVPDSsHCQVLLLP-FAEChK
K6-8 VLAPATFYAIC
V8 FrII EQCLVPDSSHC-V------AECH---ap →
[Girma et al. (1984)]
1410 1420 1430 1440 1450 1460 1470 1480 1490 1500
QQDSCHQEQVCEVIASYAHLCRTNGVCVDWRTPDFCAMSCPPSLVYNHCEHGCPRHCDGNVSSCGDHPSEGCFCPPDKVMLEGSCVPEEACTQCIGEDGV
M29C-M30-D21 SCHQEQVCEVIASYAHLCRT →
M29C-M30-D12 WRTPDFCAM
M30 SCPPSLVYNHCEHGCPRHCDG-V-s-g →
M31 LEGSCVPEEACTQCIGEDGV
M31-K5-4 LEGSCVPEEACTQCIGED →
K6-8 cont. QQDSCHQEQVCEVIASYAHLC-T-g →
K6-8-T9 TNGVCVDWR
K6-8-T14 TPDFCAMSCPPSLVY →
K6-8-T1 HCDG-VSSCGDHPSEGCFCPPDK
K3-4 VMLEGSCVPE →
M31-E11 DGV
K9-11 V-LEGSCVPEEACTQCIGEdGV
1510 1520 1530 1540 1550 1560 1570 1580 1590 1600
QHQFLEAWVPDHQPCQICTCLSGRKVNCTTQPCPTAKAPTCGLCEVARLRQNADQCCPEYECVCDPVSCDLPPVPHCERGLQPTLTNPGECRPNFTCACR
M31 cont. QHqFL-AWVP--Qp →
M31-K2-1 V-CTTQPCP-AKAP--GL-EVARLr →
M31-E19 RGLQPTLT-P →
M31-E11 cont. QHQFLE
M31-K3-2 APTCGLCEVARLRQNA-QCcP-Y →
M31-E24 AWVPDHQPCQICTCLSGRKV-CTTQPCP-AKAPT-GLC-
K3-4-T7 GLQPTLTNPGECRP-FTCACR
K3-4-T13-C11 EAWVPDHQPcQICTcL
K9-11 cont. Q--FLEA-VPd →
K3-4-T9 LRQNADQCCPEYECVCDPVSCDLpPVPHCER
1610 1620 1630 1640 1650 1660 1670 1680 1690 1700
KEECKRVSPPSCPPHRLPTLRKTQCCDEYECACNCVNSTVSCPLGYLASTATNDCGCTTTTCLPDKVCVHRSTIYPVGQFWEEGCDVCTCTDMEDAVMGL
M31-K7-1 RVSPPSCPPHRLPTLRK
M31-K4-4 TQCCDEYECACNCV-STVSCPLGYLAStAtNd →
M31-K6-2 V-VHRSTIYPVGQFW--G-DV-t →
M32 edavm
M33 GL
M31-E17 CKRV-PPSCP-HRL-LRK →
M31-E30-C10 LASTATNDGCTTTTCLPDKV-V →
K3-4-T7 cont. KEECK
K8-5-C17 LASTATNDCGCTTttCLPDk
K7-16 VCVHRSTIYPVGQFWE-G-DVCTCTDM--AVMG-
K7-16-T15 STIYPVGQFWEEGCDV-TCTDMEDAVMGL
1710 1720 1730 1740 1750 1760 1770 1780 1790 1800
RVAQCSQKPCEDSCRSGFTYVLHEGECCGRCLPSACEVVTGSPRGDSQSSWKSVGSQWASPENPCLINECVRVKEEVFIQQRNVSCPQLEVPVCPSGFQL
M33 cont. RVAQCSQKPCEDSCRSGFTYVLH-G--CG--1P →
M33-E18 VVTGSPRGD-QS →
M33-K12 SVGSQWASPENP-LINE-v--K
M33-K16 EEVFIQQ--V-CPQLEVP →
K7-16 cont. -VAQ-S →
K7-16-T10 SGFTYVLHEGECCGR
K7-16-T8 CLPSACEVVTGSPR
M33-E6 CVRVKEE
M33-E14 VPVCPSGFQL
K7-16-T15 cont. R
K11-15 SVGSQWASPENPCLINECVRVK
K9-16 EEVFIQQR-VSCPQLEVPVCPSGFQL
R11-21 GDSQSSWKSVGSQWASPENPC →

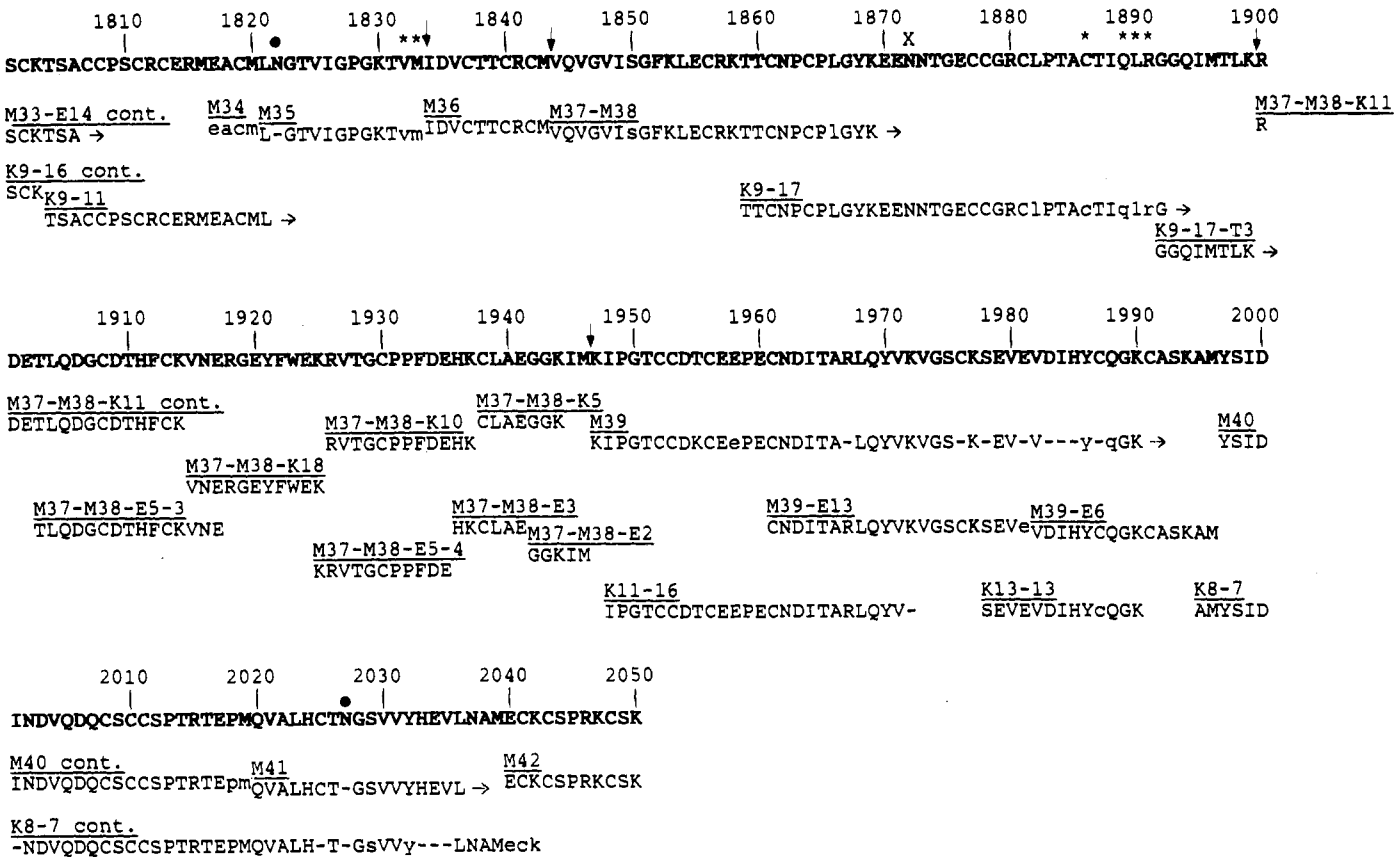


FIGURE 3: Detailed summary of the proof of sequence of human vWF. The proven sequences of specific peptides, with underlined names, are given in one-letter code below the summary sequence (bold type). Prefixes M, K, and R denote peptides generated by cleavage of the CM protein at methionyl, lysyl, and arginyl bonds, respectively. Subpeptides are identified by hyphenated suffixes, with the following code indicating the subdigesting agent: C, E, K, T, or R, enzymatic cleavage with chymotrypsin, *S. aureus* protease (at Glu), *Achromobacter* protease I (at Lys), trypsin, or trypsin of citraconylated protein (at Arg), respectively; D or W, chemical cleavage with dilute acid (at Asp) or BNPS-skatole (at Trp). Peptide sequences written in upper-case letters were proven by Edman degradation; those in lower-case letters indicate tentative identification, or deduced from amino acid compositions. Those not identified are shown by dashes or by arrows, which indicate a long unidentified sequence. Certain sequences in bold face (residues 1–10, 911–920, 1366–1393, and 2040–2050) refer to those published previously from our laboratories. Probable glycosylation sites are indicated by (●) (N-glycosylation) or by (○) (O-glycosylation) above the summary sequence. Currently, tentative residues and places where peptide overlaps are missing are also indicated by (*) and (↓) above the summary sequence. The cDNA sequence has provided data corroborating all except residues 110–525.

fragmented segments of sequence. In addition, 22 sites of N- and O-glycosylation (Figure 3) were implicated by the *lack* of identifiable phenylthiohydantoins at sites of known composition or cDNA sequence.

Cleavage at Lysyl and Arginyl Bonds. Twenty-five milligrams of [^{3}H]CM-vWF was digested in 50 mM Tris-HCl, pH 9, with 10 μg of *Achromobacter* protease I at 37 °C for 6 h and then fractionated on an HPLC sizing column equilibrated in 6 M guanidine hydrochloride–10 mM sodium phosphate buffer, pH 6 (Figure 6). Pooled fractions were further purified by RP-HPLC, and aliquots of purified peptides were subjected to amino acid analysis in order to identify methionine-containing peptides. These were numbered K1-1, K1-2, etc., denoting the order of elution from the HPLC sizing column (Figure 6) hyphenated to that from subsequent RP-HPLC columns (not shown).

Likewise, methionine-containing peptides were isolated from a tryptic digest of 25 mg of [^{14}C]methyl-Met-vWF after citraconylation of lysyl residues. These were numbered in an analogous manner as R1-1, R1-2, etc. (Figure 7). Sequenator analyses of these peptides provided 30 overlaps for the cyanogen bromide fragments as shown in Figure 2 and 3.

Amino Acid Sequence. As indicated in Figure 3, the sequence analysis of 13 cyanogen bromide fragments was not completed. Some of them (M6, M15, M19, M20, M24, M25, M31, M35, and M37) contain residues that were only tentatively identified. Others (M20, M24, M30, M31, and M38) lacked overlaps for certain subpeptides. In addition, 10 overlaps of those cyanogen bromide fragments are not proven. However, the information not provided by the direct analysis of the protein sequence was all within the carboxyl-terminal segment (residues 526–2050) of the protein that was predicted from the nucleotide sequence of the cDNA clone λHvWF3 (Sadler et al., 1985). The gap (residues 111–525) separating the two sequences predicted from the nucleotide sequences was identified by protein sequencing (Figure 3). Thus, by a combination of both cDNA and protein sequencing, the complete amino acid sequence of the 2050 residues of human vWF (Figure 8) has been established.

Discussion

Determination of the primary structure of human vWF involved two sets of data, one provided by characterization of two cDNA clones (Sadler et al., 1985) and the other by the direct analysis of the protein as described in the present data. Although each set of data is incomplete at present, the combination of the two sets of data provides the first complete amino acid sequence of the 2050 residues in the mature protein. The results indicate that proteolytic processing events must have occurred in the amino-terminal portion of a larger precursor to yield the mature, circulatory form. In addition, we have identified 22 probable glycosylation sites.

Table I

Amino Acid Compositions[a] of Cyanogen Bromide Fragments of Human von Willebrand Factor

Fragment	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10	M11
Residue No.	1-8	9-37	38-39	40-51	52-109	110-184	185-288	289-292	293-540	541	542-622
Asp/Asn (D/N)		3.8 (4)			6.1 (6)	6.3 (6)	16.0 (15)		25.7 (25)		3.5 (3)
Thr (T)		2.0 (2)			4.4 (5)	3.3 (3)	4.8 (5)	1.1 (1)	14.1 (13)		2.4 (2)
Ser (S)	1.6 (2)	1.0 (0)	1.0 (1)	1.0 (1)	2.0 (1)	3.3 (3)	8.8 (11)		15.1 (15)		8.0 (9)
Glu/Gln (E/Q)		3.5 (4)			5.5 (5)	7.8 (7)	10.1 (11)	1.0 (1)	32.5 (29)		9.1 (9)
Pro (P)	1.9 (2)	1.3 (1)		2.3 (2)	3.2 (2)	4.2 (4)	4.2 (3)		22.1 (20)		2.2 (2)
Gly (G)		1.9 (1)	0.5 (0)	3.1 (3)	4.4 (4)	9.0 (9)	6.0 (6)		13.0 (10)		3.9 (3)
Ala (A)		2.4 (2)			2.8 (3)	2.0 (1)	3.4 (3)		15.0 (16)		7.0 (7)
CMCys (C)	1.2 (1)	2.3 (4)		1.4 (3)	8.8 (9)	2.6 (4)	3.1 (4)		24.0 (28)		
Val (V)		2.7 (2)		1.3 (1)	4.8 (5)	7.2 (8)	9.2 (11)		25.0 (24)		7.1 (7)
Met[c] (M)	0.5 (1)	0.6 (1)	0.3 (1)	0.7 (1)	0.6 (1)	0.6 (1)	0.8 (1)	0.5 (1)	0.4 (1)	(1)	0.3 (1)
Ile (I)					1.9 (2)	3.0 (3)	4.8 (5)		7.0 (7)		5.3 (6)
Leu (L)	1.0 (1)	3.6 (4)		0.9 (1)	1.2 (1)	7.0 (7)	10.0 (10)		20.7 (21)		7.8 (8)
Tyr (Y)		0.9 (1)			0.9 (1)	3.2 (4)	2.2 (2)		6.3 (7)		3.2 (4)
Phe (F)					0.9 (1)	3.7 (4)	3.3 (3)		7.4 (7)		2.3 (2)
His (H)					3.0 (3)	2.0 (2)	2.6 (3)		6.9 (8)		2.0 (2)
Lys (K)		2.0 (2)			3.4 (3)	5.4 (6)	4.8 (6)	1.0 (1)	7.2 (6)		5.5 (6)
Arg (R)	1.1 (1)	1.1 (1)			5.0 (5)	3.5 (3)	3.1 (3)		9.7 (9)		8.7 (9)
Trp[d] (W)					(1)		(2)		(2)		(1)
Total	(8)	(29)	(2)	(12)	(58)	(75)	(104)	(4)	(248)	(1)	(81)

Fragment	M12	M13	M14	M15	M16	M17	M18	M19	M20	M21	M22
Residue No.	623-630	631-710	711-732	733-758	759-765	766-782	783-843	844-947	948-998	999-1022	1023-1097
Asp/Asn (D/N)		7.0 (7)	1.2 (1)	3.2 (4)		1.6 (2)	6.5 (6)	9.9 (12)	4.4 (4)	1.2 (1)	8.9 (10)
Thr (T)		2.3 (1)	1.9 (2)			1.0 (1)	2.6 (3)	3.4 (3)	2.5 (3)	0.9 (1)	2.9 (3)
Ser (S)	1.0 (1)	4.9 (4)	1.5 (2)	1.8 (2)		1.9 (2)	4.7 (5)	6.4 (8)	4.4 (5)	1.8 (2)	3.3 (5)
Glu/Gln (E/Q)	2.8 (3)	13.4 (11)	1.2 (1)	3.1 (3)	3.1 (3)	2.0 (2)	10.4 (9)	11.9 (13)	4.5 (3)	4.4 (4)	3.2 (3)
Pro (P)	1.0 (1)	8.4 (9)	1.8 (2)				2.3 (2)	9.0 (13)	3.1 (3)	1.0 (1)	4.1 (4)
Gly (G)		4.8 (3)	3.9 (4)	2.5 (2)		1.6 (1)	5.0 (5)	7.4 (8)	3.2 (2)	4.0 (4)	5.8 (6)
Ala (A)	0.9 (1)	5.7 (5)	1.0 (1)	1.9 (2)			3.0 (3)	6.0 (6)	4.0 (4)	1.9 (2)	11.3 (12)
CMCys (C)		1.3 (1)						3.2 (3)			
Val (V)		6.9 (8)	2.6 (3)	3.2 (3)	0.7 (1)	2.9 (3)	4.4 (5)	7.2 (7)	5.5 (7)	1.0 (1)	8.1 (10)
Met[c] (M)	0.5 (1)	0.4 (1)	0.6 (1)	0.5 (1)	0.6 (1)	0.6 (1)	0.4 (1)	0.3 (1)	0.5 (1)	0.5 (1)	0.5 (1)
Ile (I)		4.2 (6)		0.9 (1)	0.6 (1)	1.0 (1)	1.6 (2)	5.9 (8)	3.7 (4)	0.9 (1)	4.1 (5)
Leu (L)		8.0 (8)	2.5 (3)	2.1 (2)		1.3 (1)	5.6 (6)	9.7 (12)	5.0 (5)	2.0 (2)	5.0 (5)
Tyr (Y)		1.9 (2)				1.6 (2)	4.0 (4)	1.6 (1)	0.8 (1)	1.1 (1)	0.9 (1)
Phe (F)		2.4 (2)		2.7 (3)			2.7 (2)	2.4 (3)	2.2 (2)	1.0 (1)	1.1 (1)
His (H)		1.5 (1)				0.9 (1)	0.9 (1)	1.0 (0)	1.1 (1)		1.0 (1)
Lys (K)		6.1 (7)	1.1 (1)	1.8 (2)			1.7 (1)	2.5 (1)	6.0 (4)		2.1 (2)
Arg (R)	1.1 (1)	4.9 (4)	1.0 (1)	1.0 (1)	1.2 (1)		5.8 (6)	5.1 (4)	1.6 (1)	1.9 (2)	5.5 (6)
Trp[d] (W)								(1)	(1)		
Total	(8)	(80)	(22)	(26)	(7)	(17)	(61)	(104)	(51)	(24)	(75)

Table I (continued)

Fragment	M23	M24	M25	M26	M27	M28	M29	M30	M31	M32	M33
Residue No.	1098-1117	1118-1239	1240-1257	1258-1275	1276-1284	1285-1334	1335-1438	1439-1480	1481-1693	1694-1698	1699-1816
Asp/Asn (D/N)	1.2 (1)	13.2 (14)	1.3 (1)	2.1 (2)	1.0 (1)	6.2 (7)	7.3 (7)	6.2 (5)	19.0 (17)	0.9 (1)	6.2 (5)
Thr (T)	1.0 (1)	8.7 (10)		1.0 (1)		3.2 (4)	6.3 (9)		16.6 (21)		3.7 (3)
Ser (S)	1.8 (2)	9.0 (9)	3.1 (4)	1.1 (1)		2.0 (2)	5.1 (4)	5.3 (5)	10.1 (9)		12.7 (16)
Glu/Gln (E/Q)		16.6 (14)	2.5 (2)	1.6 (1)	1.0 (1)	5.5 (6)	14.0 (16)	2.9 (2)	26.8 (27)	1.1 (1)	15.0 (18)
Pro (P)		8.8 (8)		1.2 (1)		1.8 (2)	5.7 (6)	6.0 (6)	18.7 (22)		7.5 (9)
Gly (G)	2.0 (2)	11.0 (11)		2.7 (3)	1.2 (1)	3.9 (4)	6.6 (3)	4.6 (4)	13.8 (11)		8.9 (8)
Ala (A)		3.0 (2)	1.1 (1)		0.8 (1)	2.2 (2)	7.0 (7)		9.5 (10)	1.0 (1)	4.0 (4)
CMCys (C)	1.2 (2)	8.6 (12)				1.7 (2)	7.2 (10)	4.8 (7)	26.2 (34)		12.7 (16)
Val (V)	1.8 (2)	7.7 (9)	1.8 (2)	4.0 (5)	1.2 (2)	1.7 (1)	9.3 (10)	3.1 (3)	14.0 (15)	1.0 (1)	7.5 (11)
Met[c] (M)	0.4 (1)	0.4 (1)	0.5 (1)	0.2 (1)	0.3 (1)	0.4 (1)	0.4 (1)	0.5 (1)	0.3 (1)	0.3 (1)	0.7 (1)
Ile (I)	1.0 (1)	1.7 (2)	0.9 (1)		0.9 (1)	1.5 (2)	3.5 (3)		3.4 (3)		1.9 (2)
Leu (L)	3.0 (3)	8.6 (8)	2.0 (2)	1.0 (1)		3.7 (4)	8.6 (10)	1.4 (1)	13.0 (13)		6.0 (6)
Tyr (Y)		0.8 (1)		1.1 (1)	1.0 (1)	0.8 (1)	2.0 (2)	1.1 (1)	4.8 (4)		2.0 (1)
Phe (F)	2.0 (2)	3.1 (3)				5.0 (6)	2.5 (3)	1.0 (1)	3.5 (3)		2.9 (3)
His (H)	0.9 (1)	3.5 (4)	1.8 (2)			2.4 (3)	2.9 (4)	3.7 (4)	5.5 (5)		1.1 (1)
Lys (K)	1.0 (1)	5.3 (5)	2.0 (2)			2.0 (2)	3.2 (2)	1.1 (1)	5.8 (6)		4.5 (4)
Arg (R)	1.0 (1)	7.3 (7)		1.0 (1)		1.3 (1)	4.8 (4)	1.4 (1)	10.4 (10)		7.6 (8)
Trp[d] (W)		(2)					(3)		(2)		(2)
Total	(20)	(122)	(18)	(18)	(9)	(50)	(104)	(42)	(213)	(5)	(118)

Fragment	M34	M35	M36	M37 - M38	M39	M40	M41	M42	Hydrolysis of vWF[b]
Residue No.	1817-1820	1821-1833	1834-1843	1844-1946	1947-1996	1997-2019	2020-2039	2040-2050	1-2050
Asp/Asn (D/N)		1.0 (1)	1.0 (1)	7.6 (8)	4.6 (4)	4.0 (4)	2.0 (2)		172.6 (187)
Thr (T)		2.1 (2)	1.7 (2)	7.9 (9)	3.0 (3)	1.9 (2)	1.0 (1)		102.2 (116)
Ser (S)				2.1 (1)	3.1 (3)	2.6 (3)	1.2 (1)	1.7 (2)	131.6 (141)
Glu/Gln (E/Q)	1.0 (1)			13.9 (14)	7.2 (7)	3.7 (3)	2.2 (2)	1.1 (1)	241.2 (237)
Pro (P)		1.1 (1)		5.4 (5)	2.7 (2)	2.0 (2)		1.0 (1)	124.8 (136)
Gly (G)		3.1 (3)		11.4 (12)	4.3 (3)		1.7 (1)		146.8 (137)
Ala (A)	1.0 (1)			2.5 (2)	2.3 (3)		2.0 (2)		104 (104)
CMCys (C)	+ (1)		2.5 (3)	8.1 (11)	6.1 (7)	2.2 (3)	0.8 (1)	2.6 (3)	152.0 (169)
Val (V)		1.9 (2)	0.8 (1)	4.7 (5)	3.8 (4)	1.1 (1)	4.0 (4)		153.8 (184)
Met[c] (M)	0.6 (1)	0.7 (1)	0.6 (1)	1.4 (2)	0.7 (1)	0.6 (1)	0.4 (1)		34.8 (41)
Ile (I)		0.7 (1)	1.0 (1)	3.7 (4)	2.8 (3)	1.7 (2)			68.8 (78)
Leu (L)		1.0 (1)		6.7 (7)	2.3 (1)		2.3 (2)		150.4 (156)
Tyr (Y)				1.4 (2)	1.7 (2)	1.2 (1)	1.0 (1)		52.4 (49)
Phe (F)				3.4 (4)					55.6 (56)
His (H)				2.0 (2)	0.9 (1)		1.7 (2)		45.7 (52)
Lys (K)		0.9 (1)		4.8 (8)	6.9 (5)			2.9 (3)	84.2 (88)
Arg (R)			1.3 (1)	5.8 (6)	1.9 (1)	1.0 (1)		1.1 (1)	96.8 (101)
Trp[d] (W)				(1)					(18)
Total	(4)	(13)	(10)	(103)	(50)	(23)	(20)	(11)	(2050)

[a]Residues per fragment by amino acid analysis or, in parentheses, from the sequence (Figure 3).

[b]Hydrolysis was carried out with 6 N HCl for 24 h. The composition was calculated on the bais of 104 alanyl residues.

[c]Measured as homoserine.

[d]Not determined.

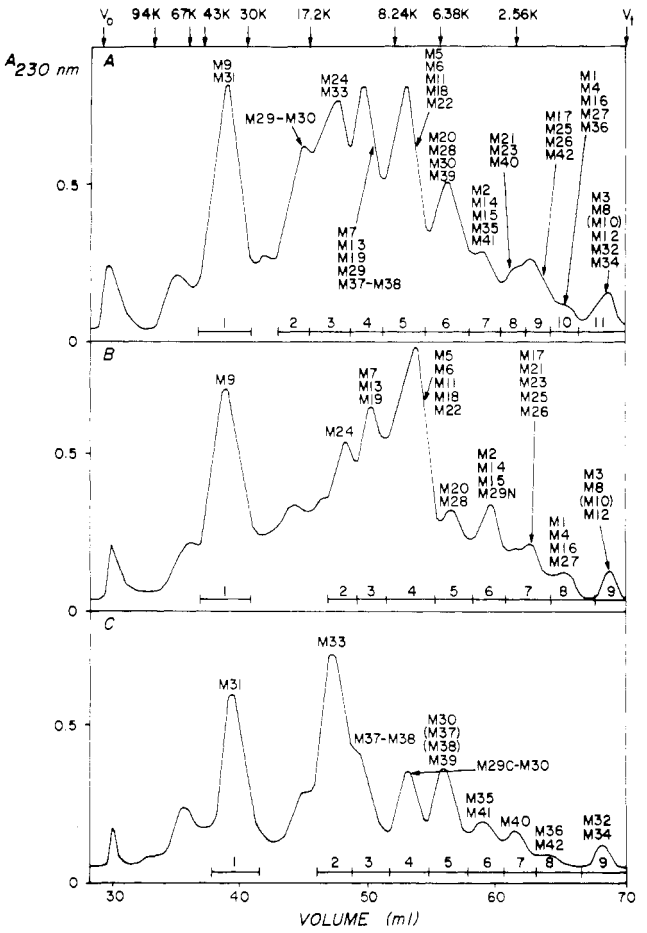


FIGURE 4: Primary separation of cyanogen bromide digests of the whole protein (A) and *S. aureus* V8 fragments III (B) and II (C). In each case, the digest (5 mg) was separated on three columns of TSK G3000SW (7.5 × 600 mm) connected in series and equilibrated in 6 M guanidine hydrochloride–10 mM potassium phosphate, pH 6.0. Peptides were monitored at 230 nm (1.0 AUFS), collected at 0.5 mL/min in 0.5-mL fractions, and pooled as shown by horizontal bars. V_0, V_t and the elution position of standard proteins or peptides of known molecular weight are shown above by vertical arrows.

The two sets of data agree in entirety where the corresponding information is available, except for residues 7 and 26. At residue 7, the presence of proline was determined by three sets of direct protein sequencing data, namely, the composition of fragment M1 and sequenator analyses of both the whole protein and peptide K13-11. At residue 26, the sequenator analysis of fragment M2 identified both Ala, and Thr at a molar ratio of about 4:1. In contrast, the nucleotide sequence of the λHvWF1 clone predicted His and Thr, respectively, at these positions (Sadler et al., 1985). This discrepancy may be due to polymorphism in the protein or to an error in DNA replication during the preparation of the cDNA library.

The polypeptide moiety of the protein (without carbohydrate) corresponds to a M_r of 225 843. Since the protein contains approximately 18.7% (w/w) carbohydrate (Chopek et al., 1986) distributed among the 22 oligosaccharide chains, the subunit molecular weight of the mature protein is approximately 278 000, which is slightly larger than previously reported from ultracentrifugation data [e.g., Legaz et al. (1973)].

It should be noted that 6.15 kilobases of mRNA would suffice to encode the mature form of human vWF (2050 residues) that circulates in blood. Recent studies of Lynch et al. (1985) and Ginsburg et al. (1985) showed that the size of mRNA identified by Northern blot analysis is approximately 8–10 kilobases, which leaves more than 2 kilobases of mRNA to encode a prepro leader sequence. Recent results of Fay et al. (1985) suggest that a 100-kilodalton plasma glycoprotein is identical with vWF antigen II, originally described by Montgomery and Zimmerman (1978), and that it represents the precursor peptide.

Several other structural features of vWF are summarized in Figure 9. There appear to be 11 asparagine- (N-) linked glycosylation sites, which are in the characteristic sequence Asn-X-Thr/Ser. Two additional Asn-X-Thr/Ser sequences (Asn-452 and -1872) are *not* glycosylated. One additional Asn (residue 384) is glycosylated in the sequence Asn-X-Cys. Glycosylation of this sequence was observed previously in protein C by Stenflo and Fernlund (1982). Six other Asn-X-Cys sequences in vWF are *not* glycosylated. It is likely that ten O-linked glycosylation sites occur at eight of the 116 threonyl and two of the 141 seryl residues. As observed in other proteins, the rules dictating the sites of O-glycosylation do not appear to be as sharply defined as those directing N-glycosylation, although in all cases proline residues are found in the vicinity of the O-glycosylation sites.

The glycosylation sites are most abundant in the amino- and carboxyl-terminal regions. In addition, eight of the O-glycosylation sites are clustered in two narrow regions, comprising residues 485–500 and 705–724. The physiological importance of the glycosylation and its clustering in vWF remains to be explored. It will be interesting to see how the sites of glycosylation correlate with the extended, asymmetric, multimeric models of vWF interpreted from electron micrographs (Slayter et al., 1985; Fowler et al., 1985).

In an attempt to suggest three-dimensional features from the amino acid sequence, the hydropathy index and the probable antigenic sites were examined. The two sets of results showed reasonable agreement for the locations of hydrophilic regions. The three most likely antigenic sites (1597–1608, 566–579, and 1114–1129) were the most hydrophilic regions, with hydropathy indexes exceeding −21. However, the hydropathy profile showed rather even distribution of hydrophilic and hydrophobic regions throughout the molecule without providing a clear indication of the regions involved in construction of the nonglobular structures envisioned in electron micrographs. The carboxyl-terminal 200 residues appear to be the most hydrophilic when searched with a wide window size (100 residues), the next being a 120-residue region around residue 1570.

Nine of the 12 N-glycosylated sites are found in hydrophilic regions, and two (468 and 1822) are found in rather hydrophobic regions. Most of the O-glycosylation sites are found in slightly hydrophobic regions. The platelet binding tetrapeptide sequence (RGDS, residues 1744–1747) is part of a very hydrophilic region (hydropathy index ca. −14) and is likely to be located on the surface of the molecule.

Browning et al. (1983) reported that human vWF was radiolabeled when endothelial cells are cultured in the presence of [^{35}S]sulfate, and the 225-kd mature protein is immunoisolated from the culture medium or cell lysate. We have not yet found the site of sulfation. This may be due to the instability of sulfated amino acid residue(s) during acid hydrolysis or Edman degradation. Alternatively, any of the ten O-glycosylation sites that we have proposed might be sulfated Thr or Ser residues.

As observed by Sadler et al. (1985) and Shelton-Inloes et al. (1986), the sequence of human vWF shows evidence of perhaps five sets of internal gene duplications, indicating that the evolution of this exceptionally long polypeptide chain has

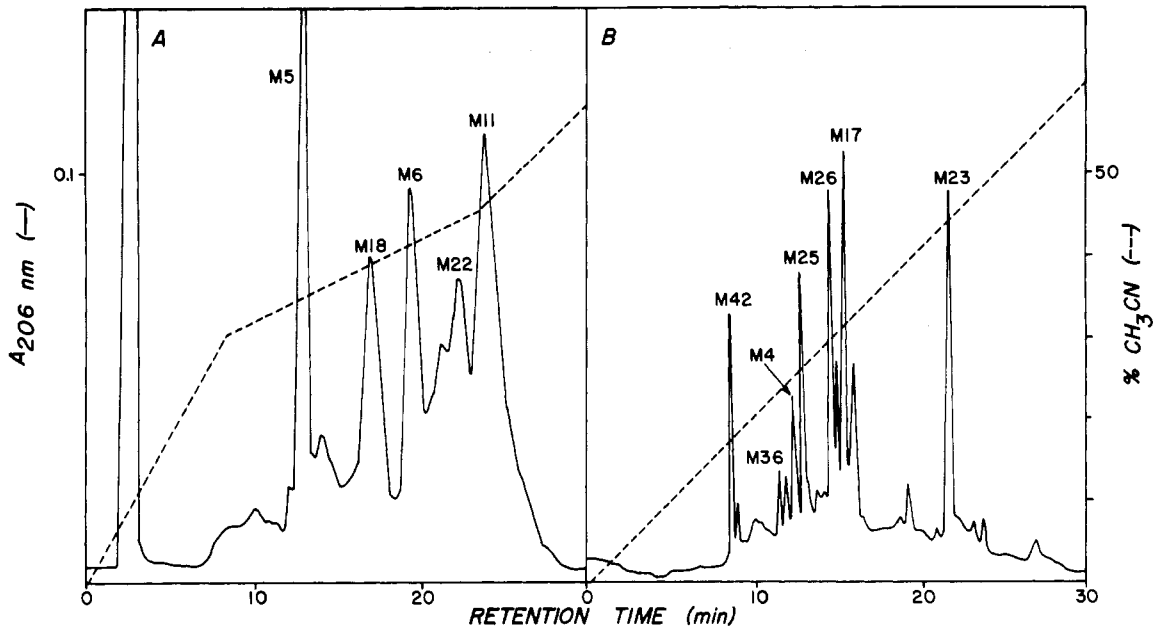


FIGURE 5: RP-HPLC separations of peptides in pooled fractions from Figure 4. (A) Separation of pool 5 (Figure 4A) on a Cosmosil 5C18P column (4.6 × 150 mm) with a TFA–acetonitrile system. (B) Separation of pool 9 (Figure 4A) on a Synchropak RP-P (C18) column (4.1 × 250 mm) with a TFA–acetonitrile system. Purified peptides are identified by the prefix M, as in Figures 2 and 3.

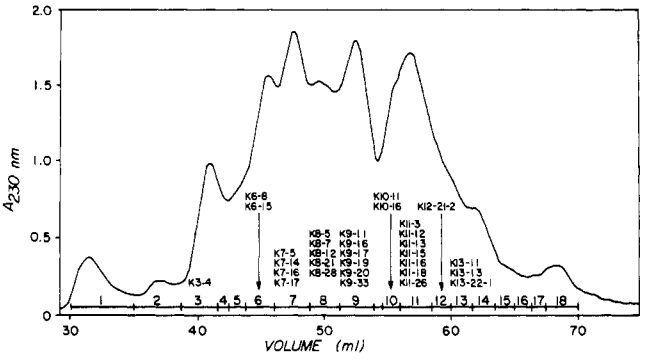


FIGURE 6: Primary separation of an *Achromobacter* protease I digest of [^{3}H]CM-vWF on the HPLC sizing column system described in Figure 4.

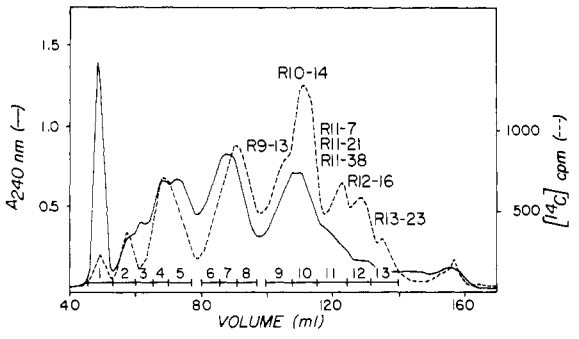


FIGURE 7: Primary separation of a 2-h tryptic digest of citraconylated [^{3}H]CM protein (25 mg) on a Sephadex G-75 superfine column (1.5 × 190 cm) equilibrated in 0.1 M NH_4HCO_3 adjusted to pH 8.8 with concentrated NH_4OH. Peptides were collected in 2-mL fractions at a flow rate of 4 mL/h. Twenty-five microliters of alternate fractions was examined for *S*-[^{14}C]methylmethionyl residues (broken line).

involved a complex pathway of divergence of duplicated regions.

vWF is a cystine-rich protein in which 169 out of 2050 residues are half-cystinyl residues (8.2%) and all of those residues are occupied in intrachain or interchain disulfide bridges. As reported previously (Girma et al., 1986), *S. aureus* V8 protease cleaves native vWF at one major and one minor site, generating three fragments (I–III), which can be separated without cleavage of disulfide bonds. Fragment I appears to be monomeric, whereas fragments II and III are each polymeric (dimeric or tetrameric). These data led to a model for the arrangement of subunits in vWF, in which it is suggested that the 270K-dalton subunits are linked by disulfide bonds that connect two carboxyl-terminal regions to each other and two amino-terminal regions to each other in head-to-head and tail-to-tail manner (Girma et al., 1986). Half-cystinyl residues are clustered and particularly abundant in fragment II, where 95 of the carboxyl-terminal 685 residues (or nearly 14% of the segment) form disulfide bonds. One-third of these residues are in Cys-X-Cys or in Cys-Cys sequences, indicating an extensive network of cross-links. Among the amino-terminal 510 residues of vWF, 56 (11%) are half-cystinyl, as indicated in Figure 9, and 14 are also in Cys-Cys or Cys-X-Cys arrangements. In contrast, there is only a single Cys residue in the 396-residue segment from Ser-510 to Arg-905. A third cluster of 14 Cys residues is found between amino acids 1109 and 1238 (within fragment I), but none appear to involve interchain disulfides. Our laboratory is currently attempting to identify the extensive network of disulfide bonds that must contribute to the ability of this unusual protein to serve its multivalent role. The pattern of disulfide cross-links should also provide a guide to the authenticity of the proposed internal homologies and their relationships to substructural binding domains.

It is likely that this very large polypeptide chain contains several substructural domains that account for its various binding functions, including its affinity for collagen, for factor VIII, and for platelets, whether induced by thrombin, ADP, or ristocetin. As yet, little is known of the location of these domains within each polypeptide chain or of any modulation of these functions by the characteristic multimeric character of the protein.

A tetrapeptide sequence, Arg-Gly-Asp-Ser, has been identified as a cell surface binding segment in fibronectin (Piersbacher & Ruoslahti, 1984a,b). The identical sequence occurs at residues 1744–1747 of human vWF as shown in Figure 8. The synthetic tetrapeptide Arg-Gly-Asp-Ser inhibits the binding of vWF to thrombin- or ADP-induced platelets (Plow et al., 1985; Haverstick et al., 1985; K. Ashida and K.

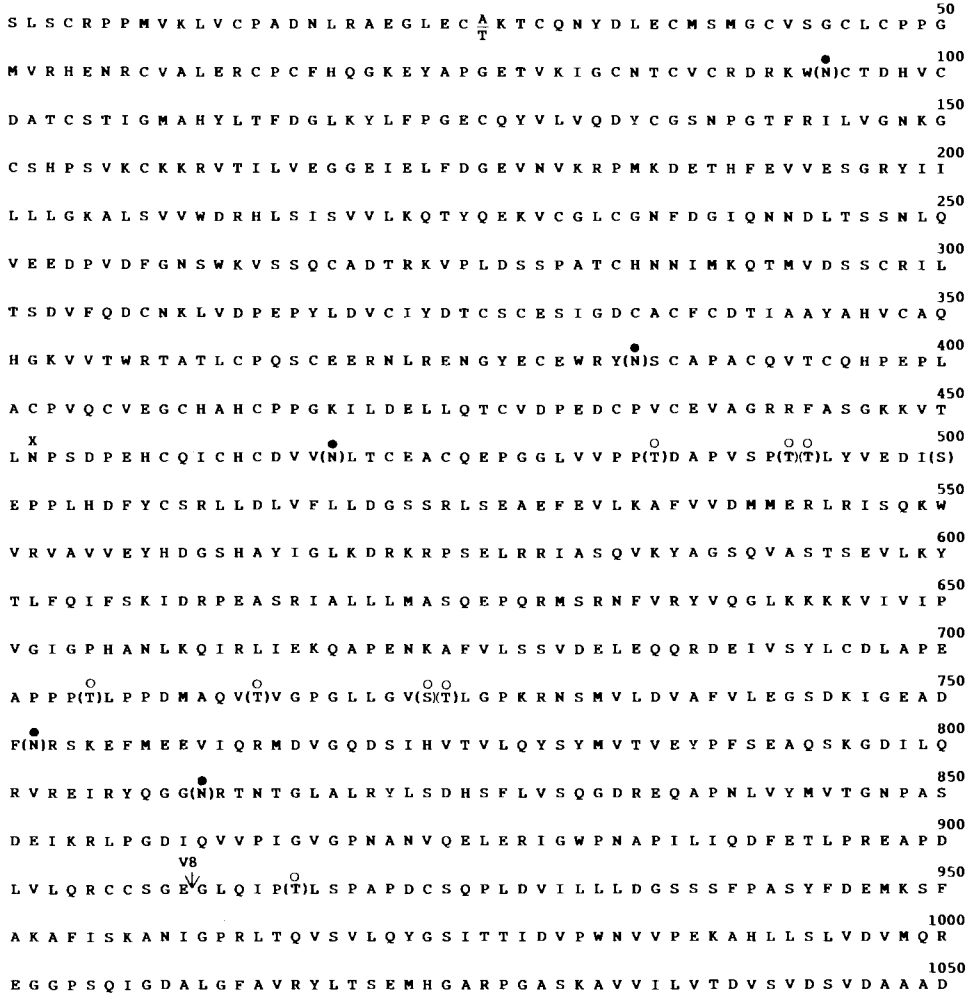


FIGURE 8: Amino acid sequence of human vWF. (●) and (○) denote sites of N- and O-glycosylation. X denotes two potential glycosylation sites that are not glycosylated. Vertical arrows indicate major (wider arrow) and minor sites of limited proteolysis by *S. aureus* V8 protease. The RGDS sequence enclosed in a rectangle (residues 1744–1747) is the probable binding site to platelets activated by ADP or thrombin.

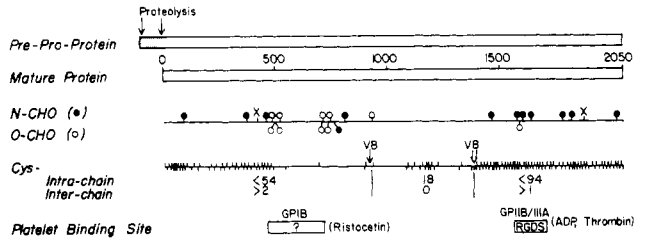


FIGURE 9: Diagrammatic view of structural features of human vWF. The amino-terminal extension includes the hypothetical sequence of a precursor form predicted by the cDNA sequence (Sadler et al., 1985). The relative orientation and clustering of oligosaccharide chains and cystine residues are indicated. On the carbohydrate line, X indicates sites where the Asn is not modified. The probable sites of binding to platelet glycoproteins GP1B and GPIIB/IIIA (platelet activating agent in parentheses) are suggested at the bottom.

Titani, unpublished data), indicating that this portion of the vWF molecule may bind to the glycoprotein IIB/IIIA on the surface of thrombin- or ADP-induced platelets. Although fragment III, generated from the amino-terminal portion of vWF by *S. aureus* V8 protease, does not contain the Arg-Gly-Asp-Ser segment, it still retains the binding activity of vWF to ristocetin-induced platelets (Girma et al., 1986).

Fujimura et al. (1985) have recently shown that a fragment of approximately 50K daltons, isolated from a limited tryptic digest of native human vWF and then reduced and alkylated, contains a domain interacting with platelet glycoprotein IB even in the absence of ristocetin. These data suggest that a binding domain for platelet glycoprotein IB may also be restricted to a local region within the molecule. The fragment, identified as residues 449–729 of the vWF molecule, is a dimer in the native form, suggesting that it contains at least one intersubunit disulfide bond.

At this time, it is not possible to place the collagen binding domain or the site of binding of factor VIIIC within the primary structure, although recent results of Fressinaud et al. (1985) show that fragment III, generated from the amino-terminal portion of vWF by *S. aureus* V8 protease, can substitute functionally for native vWF in supporting platelet adhesion to collagen.

ACKNOWLEDGMENTS

We are grateful for the suggestions and encouragement of Drs. Earl W. Davie and Hans Neurath, for the preparative work of Lee Hendrickson, for excellent technical assistance by Maria Harrylock and Scott Hormel, for the help of Ben Hodge, Meta Zonneyville, and Janine C. Harrison, and for the patience and skill of Mary Woods in the manuscript preparation.

Registry No. vWF, 9001-27-8; blood coagulation factor VIII (human von Willebrand factor reduced), 101629-40-7.

REFERENCES

Bidlingmeyer, B. A., Cohen, S. A., & Tarvin, T. L. (1984) *J. Chromatogr. 336*, 93–104.

Browning, P. J., Ling, E. H., Zimmerman, T. S., & Lynch, D. C. (1983) *Blood 62*, 281.

Chopek, M. W., Girma, J.-P., Fujikawa, K., Davie, E. W., & Titani, K. (1986) *Biochemistry* (first paper of four in this issue).

Counts, R. B., Paskell, S. L., & Elgee, S. K. (1978) *J. Clin. Invest. 62*, 702–709.

Ericsson, L. H., Wade, R. D., Gagnon, J., McDonald, R. M., Granberg, R. R., & Walsh, K. A. (1977) in *Solid Phase Methods in Protein Sequence Analysis* (Previero, A., & Coletti-Previero, M. A., Eds.) p 137, Elsevier/North-Holland, Amsterdam.

Fay, P. J., Kawai, Y., Wagner, D. D., Ginsburg, D., Bonthron, D., Ohlsson-Wilhelm, B. M., Chavin, S. I., Abraham, G. N., Hardin, R. I., Orkin, S. H., Montgomery, R. R., & Marder, V. J. (1985) *Blood 66*(Suppl. 1), Abstr. 1219.

Fowler, W. E., Fretto, L. J., Hamilton, K. K., Erickson, H. P., & McKee, P. A. (1985) *J. Clin. Invest. 76*, 1491–1500.

Fressinaud, E., Sakariassen, K. S., Girma, J.-P., Meyer, D., & Baumgartner, H. R. (1985) *Blood 66*(Suppl. 1), Abstr. 1220.

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.

Ginsburg, D., Handin, R. I., Bonthron, D. T., & 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. 2), Abstr. 836.

Girma, J.-P., Chopek, M. W., Titani, K., & Davie, E. W. (1986) *Biochemistry* (second paper of four in this issue).

Glajch, J. L., Gluckman, J. C., Charikofsky, J. G., Minor, J. M., & Kirkland, J. J. (1985) *J. Chromatogr. 318*, 23–39.

Haverstick, D. M., Cowan, J. F., Yamada, K. M., & Santoro, S. A. (1985) *Blood 66*, 946–952.

Hewick, R. M., Hunkapiller, M. W., Hood, L. E., & Dreyer, W. J. (1981) *J. Biol. Chem. 256*, 7990–7997.

Hopp, T. P., & Woods, K. R. (1981) *Proc. Natl. Acad. Sci. U.S.A. 78*, 3824–3828.

Hovig, T., & Stormoken, H. (1974) *Acta Pathol. Microbiol. Scand., Suppl. 248*, 105–122.

Hoyer, L. W. (1982) in *The Hemophilias* (Bloom, A. L., Ed.) pp 106–121, Churchill-Livingstone, Edinburgh.

Hoyer, L. W., & Shainoff, J. R. (1980) *Blood 55*, 1056–1059.

Hunkapiller, M. W., & Hood, L. E. (1983) *Methods Enzymol. 91*, 486–493.

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.

Jorgensen, L., & Borchgreven, C. F. (1964) *Acta Pathol. Microbiol. Scand. 60*, 55–82.

Kinoshita, S., Harrison, J., Lazerson, J., & Abildgaard, C. F. (1984) *Blood 63*, 1369–1371.

Kyte, J., & Doolittle, R. F. (1982) *J. Mol. Biol. 157*, 105–132.

Legaz, M. E., Schmer, G., Counts, R. B., & Davie, E. W. (1973) *J. Biol. Chem. 248*, 3946–3955.

Ling, E. H., Browning, P. H., Zimmerman, T. S., & Lynch, D. C. (1984) *Circulation 70*(Suppl. 2), Abstr. 833.

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., Morin, M. J., Ling, E. H., & Livingston, D. M. (1985) *Cell (Cambridge, Mass.) 41*, 49–56.

Masaki, T., Tanabe, M., Nakamura, K., & Soejima, M. (1981) *Biochim. Biophys. Acta 660*, 44–50.

Meyer, D., Obert, B., Pietu, G., Lavergne, J. M., & Zimmerman, T. S. (1980) *J. Lab. Clin. Med. 95*, 590–602.

Montgomery, R. R., & Zimmerman, T. S. (1978) *J. Clin. Invest. 61*, 1498–1507.

Nachman, R. L., Levine, R., & Jaffe, E. A. (1977) *J. Clin. Invest. 60*, 914–921.

Olson, J. D., Brockway, W. J., Fass, D. N., Bowie, E. J. W., & Mann, K. G. (1977) *J. Lab. Clin. Med. 89*, 1278–1294.

Perret, B. A., Furlan, M., & Beck, E. A. (1979) *Biochim. Biophys. Acta 578*, 164–174.
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.
Plow, E., Pierschbacher, M. D., Ruoslahti, E., Marguerie, G., & Ginsburg, M. (1985) *Proc. Natl. Acad. Sci. U.S.A. 82*, 8057–8061.
Ruggeri, Z. M., & Zimmerman, T. S. (1980) *J. Clin. Invest. 65*, 1318–1325.
Ruggeri, Z. M., Nilsson, I. M., Lombardi, R., Holmberg, L., & Zimmerman, T. S. (1982) *J. Clin. Invest. 65*, 1318–1325.
Sadler, J. E., Shelton-Inlos, B. B., Sorace, J. M., Harlan, J. M., Titani, K., & Davie, E. W. (1985) *Proc. Natl. Acad. Sci. U.S.A. 82*, 6394–6398.
Shapiro, G. A., Andersen, J. C., Pizzo, S. V., & McKee, P. A. (1973) *J. Clin. Invest. 52*, 2198–2210.
Shelton-Inloes, B. B., Titani, K., & Sadler, J. E. (1986) *Biochemistry* (third paper of four in this issue).
Slayter, H., Loscalzo, J., Bockenstedt, P., & Handin, R. I. (1985) *J. Biol. Chem. 260*, 8559–8563.
Stenflo, J., & Fernlund, P. (1982) *J. Biol. Chem. 257*, 12180–12190.
Takio, K., Smith, S. B., Walsh, K. A., Krebs, E. G., & Titani, K. (1984) *J. Biol. Chem. 258*, 5531–5536.
Titani, K., Sasagawa, T., Ericsson, L. H., Kumar, S., Smith, S. B., Krebs, E. G., & Walsh, K. A. (1984) *Biochemistry 23*, 4193–4199.
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.

Stabilization of Collagen Fibrils by Hydroxyproline†

George Némethy and Harold A. Scheraga*
Baker Laboratory of Chemistry, Cornell University, Ithaca, New York 14853-1301
Received November 7, 1985; Revised Manuscript Received January 27, 1986

ABSTRACT: The substitution of hydroxyproline for proline in position Y of the repeating Gly-X-Y tripeptide sequence of collagen-like poly(tripeptide)s (i.e., in the position in which Hyp occurs naturally) is predicted to enhance the stability of aggregates of triple helices, while the substitution of Hyp in position X (where no Hyp occurs naturally) is predicted to decrease the stability of aggregates. Earlier conformational energy computations have indicated that two triple helices composed of poly(Gly-Pro-Pro) polypeptide chains pack preferentially with a nearly parallel orientation of the helix axes [Nemethy, G., & Scheraga, H. A. (1984) *Biopolymers 23*, 2781–2799]. Conformational energy computations reported here indicate that the same packing arrangement is preferred for the packing of two poly(Gly-Pro-Hyp) triple helices. The OH groups of the Hyp residues can be accommodated in the space between the two packed triple helices without any steric hindrance. They actually contribute about 1.9 kcal/mol per Gly-Pro-Hyp tripeptide to the packing energy, as a result of the formation of weak hydrogen bonds and other favorable noncovalent interatomic interactions. On the other hand, the substitution of Hyp in position X weakens the packing by about 1.7 kcal/mol per Gly-Hyp-Pro tripeptide. Numerous published experimental studies have established that Hyp in position Y stabilizes an isolated triple helix relative to dissociated random coils, while Hyp in position X has the opposite effect. We propose that Hyp in position Y also enhances the stability of the *assembly* of collagen into microfibrils while, in position X, it decreases this stability.

We demonstrate that hydroxyproline plays an important role in the stabilization not only of the triple-helical collagen molecule but also of its assembly into microfibrils. The 4-hydroxyprolyl residue occurs frequently in the Y position of the Gly-X-Y repeating tripeptide of vertebrate collagens, but never in the X position. For example, in the best-characterized structure, bovine skin type I collagen, Hyp is found 112 times in the 337 triple-helical Gly-X-Y tripeptides of the α1(I) chain (Bornstein & Traub, 1979). Its stabilizing effect on the triple helix has been recognized for many years. The presence of Hyp in position Y raises the melting temperature, i.e., the temperature of the triple helix-random coil conversion, in synthetic poly(Gly-Pro-Hyp) as compared to poly(Gly-Pro-Pro) and in collagens with increasing amounts of Hyp (Kobayashi et al., 1970; Berg & Prockop, 1973; Jimenez et al., 1973; Rosenbloom et al., 1973; Sakakibara et al., 1973; Ward & Mason, 1973; Fessler & Fessler, 1974; Burjanadze, 1982). On the other hand, substitution of Hyp for Pro in position X decreases the melting temperature (Inouye et al., 1982). It has been shown, however, that no hydrogen bonds can be formed directly between the hydroxyl group of Hyp in either position X or Y and any backbone groups of the same triple helix (Ramachandran et al., 1973; Traub, 1975; Miller et al., 1980). Thus, the stabilizing effect of Hyp on the triple helix must come from some other source. It has been proposed that it is due to interactions with the solvent. Specific hydrogen bonding involving water molecules has been proposed in several models (Ramachandran et al., 1973; Traub, 1974; Bansal et al., 1975; Privalov et al., 1979; Suzuki et al., 1980).

In most studies cited, the thermal stabilities of *isolated triple helices* of poly(tripeptides)s were investigated. Very little information is available on the effect of hydroxyproline on the

† This work was supported by research grants from the National Institute of General Medical Sciences (GM-14312) and the National Institute on Aging (AG-00322) of the National Institutes of Health, U.S. Public Health Service, and from the National Science Foundation (DMB84-01811).

0006-2960/86/0425-3184$01.50/0