Proc. Natl. Acad. Sci. USA
Vol. 81, pp. 3640–3644, June 1984
Biochemistry

Complete amino acid sequence of human plasma β_2-glycoprotein I

(apolipoprotein H/activated protein C-binding protein/glucosamine oligosaccharide/internal homology/gene evolution)

JAY LOZIER*, NOBUHIRO TAKAHASHI, AND FRANK W. PUTNAM†

Department of Biology, Indiana University, Bloomington, IN 47405

Contributed by Frank W. Putnam, February 22, 1984

ABSTRACT **We have determined the complete amino acid sequence of β_2-glycoprotein I (M_r, ≈50,000), a human plasma protein that is associated with lipids and binds to platelets but whose function is not yet known. The protein consists of 326 amino acids and has five attached glucosamine-containing oligosaccharides. The protein is rich in cysteine and proline, and the sequence is notable for the frequent occurrence of Cys-Pro linkages at regular intervals. Computerized analysis of the sequence reveals five consecutive homologous segments in which cysteine, proline, and tryptophan appear to be highly conserved. This suggests that β_2-glycoprotein I may have evolved by repeated duplications of a gene coding for a 60-amino acid segment of protein.**

β_2-Glycoprotein I (β_2I; M_r, ≈50,000) is one of several human plasma proteins that have been highly purified, crystallized, and characterized, but whose physiological function remains unknown (1, 2). However, β_2I is associated with lipoproteins, binds to platelets, interacts with heparin, and may be involved in blood coagulation. As part of a program for investigation of such plasma proteins, we selected β_2I for structural study because, compared to other human plasma proteins, it has an unusually high content of cysteine (6.2%) and proline (8.3%) (calculated as g of amino acid residue per g of polypeptide). We have developed methods for purification of glycopeptides by high-performance liquid chromatography (HPLC) (3), so β_2I was also of interest to us because of its high carbohydrate content (≈19%) (1, 4).

After its discovery (4), various genetic and epidemiologic studies were done by Cleve (5) and Cleve and Rittner (6) to determine what role β_2I might play in normal metabolism and disease states; however, no function could be assigned. Other investigators have suggested that β_2I may be involved in lipoprotein metabolism. The protein has been identified as a constituent of chylomicrons, very low density lipoproteins, and high density lipoproteins (7), and it precipitates triglyceride-rich lipoproteins in the presence of anionic detergents (8). In fact, Nakaya *et al.* (9) have designated β_2I as "apolipoprotein H," proposing that it acts as an activator of lipoprotein lipase. In other studies, Schousboe (10, 11) has presented evidence that β_2I specifically binds to platelet membranes and modulates the activity of adenylate cyclase. β_2I is precipitated by heparin and transports heparin in agar gel electrophoresis (12). However, the above research has not yet provided a clear determination of the role of β_2I in normal human metabolism.

Structural studies on the protein have thus far been limited to gross characterization (13) or have been cited only in abstracts (14, 15). A preliminary report of the far-ultraviolet circular dichroism spectrum of β_2I indicates that the protein consists mainly of β sheets and random coils, with little α-helical character to its secondary structure (15). As the first phase of a detailed structural study of this protein, we have determined its complete amino acid sequence, the binding sites for the five glucosamine-containing oligosaccharides, and part of the disulfide bond structure.

MATERIALS AND METHODS

Materials. β_2I, crystallized after purification by rivanol and ammonium sulfate precipitation (16), was obtained from Behringwerke Laboratories (Marburg/Lahn, F.R.G.). The protein was judged to be pure by immunodiffusion, NaDodSO$_4$/polyacrylamide gel electrophoresis, and automated sequence analysis of the intact protein.

Methods. Prior to cleavage, the protein was reduced and alkylated with iodoacetic acid by the method of Crestfield *et al.* (17), except that dithiothreitol (Sigma) was used as a reducing agent.

CNBr (Eastman) was used to cleave the protein at methionine residues. Cleavage was done at a ratio of 200 mol of CNBr per mol of methionine in 70% formic acid for 48 hr at room temperature. Enzymatic digestion of the protein or of CNBr fragments was done using trypsin, chymotrypsin, *Staphylococcus aureus* V-8 protease, or thermolysin. Procedures for the purification of the CNBr fragments and peptides by gel filtration and HPLC have been reported, as were methods for amino acid analysis, carbohydrate analysis, and automated sequence determination with a Beckman 890C sequencer (18). The specific application of these procedures to the structural study of human β_2I is described in detail by Lozier (19), who also presents the data for the documentation of the proof of the complete sequence determination of this protein.

Computer Analysis of Sequence Data. The computer program SEARCH (20) was used to compare the sequence of β_2I with the sequence data base of the *Atlas of Protein Sequence and Structure* (updated August 15, 1983; ref. 21). The computer program RELATE (20) was used to evaluate regions of homology within the protein.

RESULTS AND DISCUSSION

Strategy for Amino Acid Sequence Analysis. Carbohydrate-rich proteins often present serious problems for structural study because of the difficulty of purifying glycopeptides and other large fragments that tend to interact, aggregate, and contaminate each other. Indeed, this is the reason that a preliminary structural study of β_2I (14) was discontinued. However, HPLC is rapidly displacing traditional procedures for the purification of peptides because of its capacity to separate large fragments suitable for sequence analysis, as well as small peptides (22). Recently, we demonstrated the purification by HPLC of all five CNBr fragments of β_2I, even though these ranged from 33 to 119 amino acids long, and although one was multiply glycosylated and was present both as a monomer and dimer (18). Therefore, the strategy

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

Abbreviations: β_2I, β_2-glycoprotein I; APC, activated protein C.

*Present address: Indiana University School of Medicine, Indianapolis, IN 46223.
†To whom reprint requests should be addressed.

for sequence analysis for β_2I was based on four considerations: (*i*) our ability to purify large polypeptide fragments and glycopeptides by use of HPLC after initial gel filtration (3, 18, 23); (*ii*) the fact that extended automated sequence analysis of these fragments was possible because of their purity; (*iii*) the rapid separation by HPLC of many smaller peptides obtained by enzymatic digestion of the whole protein or of CNBr fragments; (*iv*) identification of these peptides by amino acid composition, followed by sequence analysis of only those needed to complete the structure of CNBr fragments and to overlap them.

Determination of the Complete Amino Acid Sequence. The five CNBr fragments designated CB 1 to CB 5 in the structural model of β_2I given in Fig. 1 were separated initially by gel chromatography followed by HPLC and were characterized as described by Lozier *et al.* (18). Automated sequence analysis of these fragments was successfully performed as follows: CB 1, residues 1–40; CB 2, residues 43–72; CB 3, residues 162–193; CB 4, residues 239–268; CB 5, residues 272–308 (for numbering, see Fig. 2). The three smaller fragments (CB 1, CB 4, and CB 5) did not contain carbohydrate. Purification by HPLC of enzymatic subpeptides of CB 1, CB 4, and CB 5 followed by automated analysis gave the complete amino acid sequence of the three fragments and showed that CB 1 was amino-terminal and CB 5 was carboxyl-terminal (Fig. 2). The largest CNBr fragment (CB 2) contained 119 residues but had only one GlcN oligosaccharide attached. The primary structure was established by sequence analysis of selected enzymatic peptides of the fragment and of the intact protein. The remaining CNBr fragment (CB 3) presented the greatest difficulty because of the presence of three GlcN oligosaccharides and its tendency to aggregate. A series of enzymatic digests were made that yielded peptides needed to complete the structure. The overlapping of the CNBr fragments was accomplished by isolation and sequence analysis of selected peptides purified by HPLC from an *S. aureus* protease digest of the intact reduced and carboxymethylated protein.

By use of the strategy described above, an unambiguous amino acid sequence was established for β_2I. Each residue was validated independently in at least two peptides. In a few cases, some positions were determined with the aid of data from preliminary sequence analysis in this laboratory (14). Only the peptides essential to prove the structure are given in Fig. 2. All peptides obtained by use of HPLC fit into the structure shown. Complete data for each peptide, including amino acid composition, yield at each step, and repetitive yield, are tabulated by Lozier (ref. 19; available on request to our laboratory and also available on microfilm from University Microfilms International, 300 North Zeeb Road, Ann Arbor, MI 48106).

Polypeptide Chain Structure and Molecular Weight. β_2I consists of a single polypeptide chain containing 326 amino acid residues. The amino acid composition calculated from sequence analysis (Table 1) corresponds closely to that obtained by amino acid analysis of the protein (1). The most abundant amino acid is proline (31 residues per molecule, or approximately every 10th residue). Thus, excluding collagen and related molecules, β_2I has one of the most proline-rich structures of any eukaryotic protein. Half-cystine is the fifth most abundant amino acid (22 residues), and most of this appears to be involved in disulfide bonds. These facts, coupled with the high content of glycine (23 residues), suggest the occurrence of frequent β-turns.

The molecular weight calculated from the amino acid sequence of the unmodified polypeptide chain is 36,281. This is considerably less than the M_r of 54,200 that we estimated by $NaDodSO_4$/polyacrylamide gel electrophoresis in the presence of mercaptoethanol, and it also is less than the M_r that has been reported from sedimentation studies (40,000–48,000) (1, 13, 15). However, the high carbohydrate content of β_2I greatly affects the M_r estimated by both approaches. The five GlcN oligosaccharides are probably divided into at least two types (complex and mannose-rich), and each type is likely to be heterogeneous in structure. However, if a typical M_r of 2500 is assigned to each oligosaccharide, the M_r of β_2I will approximate 48,000–50,000.

Sequence Homology to Other Proteins. A comparison of 60-residue segments of β_2I with all known sequences in the computerized data base of the *Atlas of Protein Sequence and Structure* (19, 20) showed the protein to be unique. No other proteins in the data base exhibited significant homology over long segments. However, on checking the literature further, we noted that the amino-terminal sequence of β_2I matches exactly the amino-terminal 20 residues of a human serum protein capable of binding to human activated protein C (APC) that was isolated and named "APC-binding protein" or "APC inhibitor" by Canfield and Kisiel (24). The protein (M_r, 54,000) was postulated to bind the activated form of protein C of the blood coagulation system and to modulate coagulation in some way. However, the authors were not able to demonstrate any such activity in the APC-binding protein after purification. The identity between β_2I and the APC-binding protein seems certain, but the function of this protein remains obscure. It is interesting that the purification of APC-binding protein made use of heparin-affinity chro-

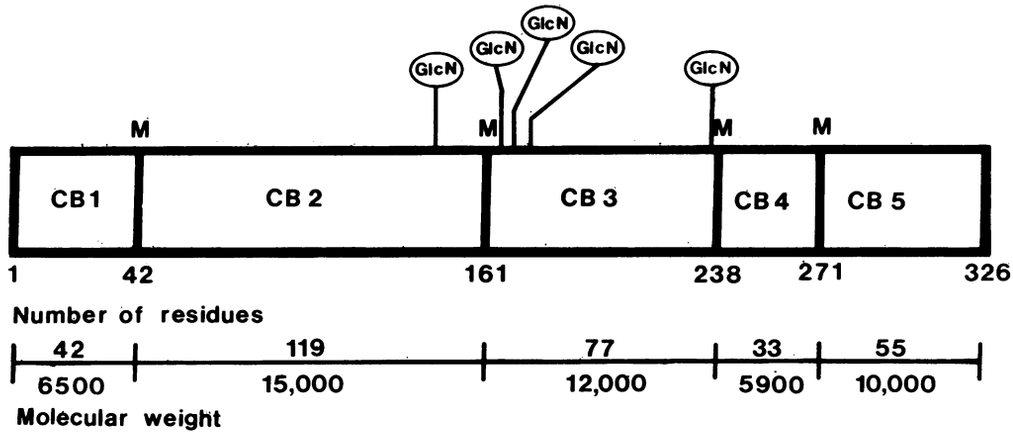

FIG. 1. Structural model of β_2I. The figure depicts the five CNBr fragments of the protein (CB 1–5) with the number of amino acids in each fragment and the approximate molecular weight of each fragment appearing below. The five glucosamine-containing oligosaccharide attachment sites are designated GlcN, and the four methionines are designated M.

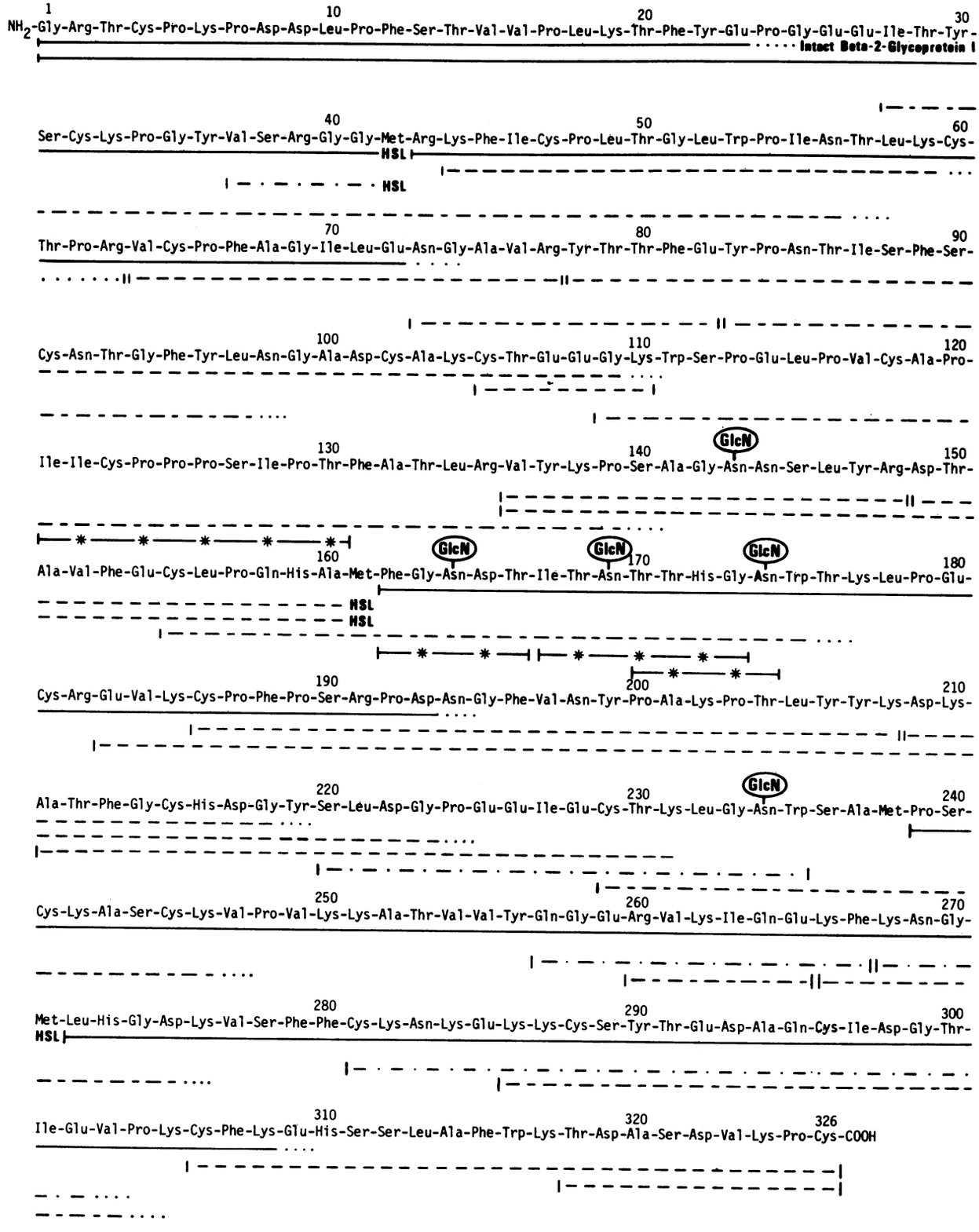

FIG. 2. Complete amino acid sequence of human plasma β_2I. The sequence is shown along with all peptides necessary for the proof of sequence. Peptides are designated by the following code:——, CNBr fragments or intact protein;— — —, tryptic peptides of CNBr fragments or intact protein;· — · — ·, chymotryptic peptides;. — — — — , *S. aureus* V-8 protease peptides;—*—, thermolysin peptides. Homoserine lactone is abbreviated HSL. Attachment sites for glucosamine-containing oligosaccharides are designated by GlcN. Detailed information about the peptides and documentation of the sequence are given in ref. 19.

matography; as mentioned before, β_2I has been shown to bind heparin (12). It may be that β_2I interacts with heparin as part of its physiological function.

Internal Homology of β_2I. This protein is unusually rich in proline (8.3%) and relatively rich in cystine (6.2%), which normally appear in proteins at frequencies of ≈5% and ≈2%,

Table 1. Amino acid composition of human β_2I based on the complete sequence determination

Amino acid	No. of residues	Amino acid	No. of residues
Aspartic acid	14	Valine	18
Asparagine	15	Methionine	4
Threonine	27	Isoleucine	13
Serine	19	Leucine	17
Glutamic acid	20	Tyrosine	14
Glutamine	4	Phenylalanine	18
Proline	31	Lysine	30
Glycine	23	Histidine	5
Alanine	17	Arginine	10
Half-cystine	22	Tryptophan	5

M_r of unmodified polypeptide chain in 36,281; number of residues is 326. Asn-143, Asn-164, Asn-169, Asn-174, and Asn-234 bind GlcN oligosaccharides.

respectively (25). In β_2I there is also a tendency for these two residues to appear together in Cys-Pro linkage, as at residues 4-5, 65-66, 123-124, and 186-187. The appearance of the Cys-Pro linkage at regular intervals of ≈60 residues raises the possibility of internal homology in the protein. Fig. 3 shows one alignment of the sequence in which consecutive groups of 60 residues are compared for homology. The figure clearly shows the conservation of cysteine-proline combinations, as well as of cysteines not associated with proline residues. Tryptophan, which is the least abundant of the 20 common amino acids in most proteins, is conserved at positions 53, 111, 175, and 235. In all, 18 of 22 cysteines are conserved and 4 of 5 tryptophan residues are conserved in this alignment. Two sets of glycines are also conserved: those homologous to glycine-40 and those homologous to glycine-51.

Further evidence for the existence of internal homology comes from the results of a computer comparison of all possible pairs of 30-residue segments of β_2I for homology, using the program RELATE and the unitary matrix (19, 20). The results of the survey show that the two most-related 30-residue segments begin at tyrosine-83 and tyrosine-207, which are separated by 124 amino acids. The 100 most related 30-residue segments of β_2I are all separated by ≈60 amino acids or some small multiple of 60. It seems that β_2I is constructed of homologous units of about 60 amino acids in which cysteines, prolines, and tryptophans are highly conserved.

Disulfide Bonding Pattern of β_2I. By thermolysin digestion of unmodified β_2I and HPLC purification of the resulting peptides, we have isolated several disulfide-linked peptides. Analysis of these peptides indicates linkage of Cys-4 and Cys-47, Cys-32 and Cys-60, Cys-91 and Cys-118, Cys-155 and Cys-181, Cys-186 and Cys-229, and Cys-281 and Cys-288. This corresponds to the pattern of internal homology presented earlier (Fig. 3) in that certain homologous sets of cysteines are linked to other homologous sets of cysteines.

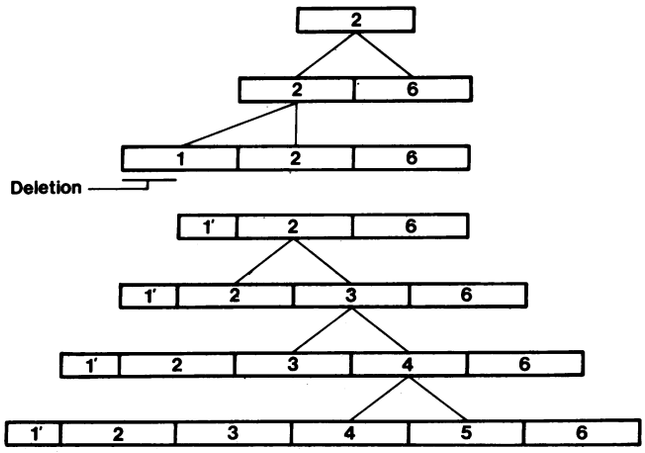

FIG. 4. Proposed model for evolution of β_2I based on computer analysis of intrasequence and intersegment homology. In this model, the primordial gene segment (corresponding to segment 2 in the sequence of Fig. 3) undergoes a tandem duplication. This is followed by a second duplication of segment 2 to yield the new segment 1, which later suffers a deletion to yield 1'. Subsequent tandem duplications as outlined yield a model with the truncated amino-terminal segment 1' and five contiguous 60-residue homologous segments designated 2–6. These correspond to the regions of sequence homology aligned in Fig. 3. Although the order of the proposed gene duplications is based on computer analysis of the amino acid sequence, the timing of the partial deletion of segment 1 is arbitrary.

Model for Evolution of β_2I. In an attempt to construct a model for the evolution of β_2I, the homologous segments were compared with one another by the computer program ALIGN, using both the unitary matrix and the mutation data matrix (19, 20). The alignment scores indicated that segment 6 is the least related to the others and thus suggested that it arose by an early duplication. The intersegment alignment scores further indicated that segment 1 had undergone a deletion to yield 1' and that subsequent duplication followed the pattern illustrated in Fig. 4. This model is based on the assumption that the most related segments have arisen from the most recent duplication events. A precedent for repeated tandem duplication is to be found in a number of plasma proteins, most notably the immunoglobulins (26). However, there is no information yet on the genomic structure of human β_2I, whereas in immunoglobulin genes the exons accord precisely with the homologous domains earlier postulated from a similar intrasequence comparison (27).

Number and Location of the Oligosaccharides. Although we did not detect galactosamine in β_2I by use of the amino acid analyzer under conditions in which both GalN and GlcN are determined, we have identified five sites at which GlcN-containing oligosaccharides are attached. All are linked to asparagine residues in the carbohydrate acceptor sequence Asn-

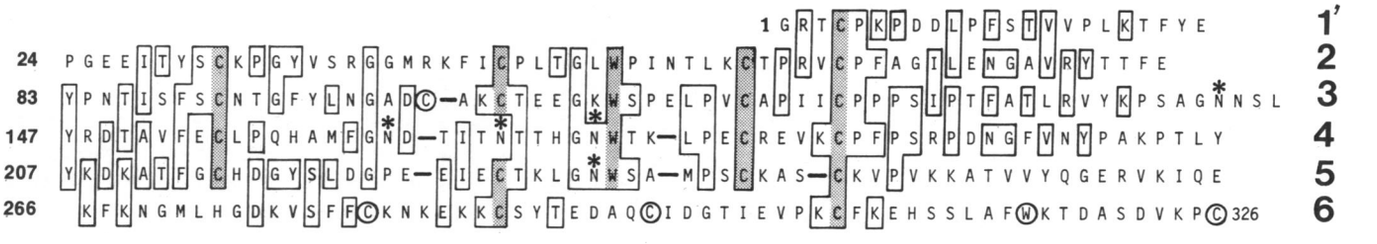

FIG. 3. Internal homology of β_2I. The sequence is aligned to demonstrate the proposed internal homology of the protein, using the one-letter code for amino acids (20). Residues that appear in equivalent positions in each segment of the protein are enclosed. Conserved cysteines and tryptophans are shaded; nonconserved cysteines and tryptophans are circled. The attachment sites for glucosamine-containing oligosaccharides are denoted by asterisks. The homologous segments are numbered 1' to 6 on the right. The position in the sequence of β_2I of the first amino acid in each segment is shown on the left (see Fig. 2).

X-Ser/Thr (28). The oligosaccharides are attached to asparagine residues at positions 143, 164, 169, 174, and 234. Two of these (at asparagine-174 and asparagine-234) are unusual in that tryptophan occupies the middle position of the acceptor sequence; it is noteworthy that these two tryptophan residues appear in homologous positions within the protein. The acceptor sequence Asn-Trp-Ser/Thr is exceedingly rare. In a search of the data base of the *Atlas of Protein Sequence and Structure*, we found only seven instances of the sequence Asn-Trp-Ser. Of these, only three were in extracellular eukaryotic proteins that could have oligosaccharides attached, and none have been shown to be so modified. Of the eight instances of the sequence Asn-Trp-Thr in the data base, only two were in extracellular proteins of eukaryotic origin, but the sequences had been inferred from DNA sequences and the question of oligosaccharide attachment is unanswered. The demonstration of the glycosylation of such acceptor sequences in β_2I establishes that glycosylation is possible for Asn-X-Thr/Ser sequences containing tryptophan as the middle residue.

We thank P. H. Davidson, S. A., Dorwin, J. A. Dwulet, K. L. Huss, J. M. Madison, and Y. Takahashi for their excellent technical assistance and Y.-S. V. Liu and F. E. Dwulet for preliminary structural studies. This work was supported by Grant AM 19221 from the National Institutes of Health.

1. Heimburger, N., Heide, K., Haupt, H. & Schultze, H. E. (1964) *Clin. Chim. Acta* **10,** 293–307.
2. Schwick, H. G. & Haupt, H. (1984) in *The Plasma Proteins*, ed. Putnam, F. W. (Academic, New York), 2nd Ed., Vol. 4, in press.
3. Tetaert, D., Takahashi, N. & Putnam, F. W. (1982) *Anal. Biochem.* **123,** 430–437.
4. Schultze, H. E., Heide, H. & Haupt, H. (1961) *Naturwissenschaften* **48,** 719.
5. Cleve, H. (1968) *Humangenetik* **5,** 294–304.
6. Cleve, H. & Rittner, C. (1969) *Humangenetik* **7,** 93–97.
7. Polz, E. & Kostner, G. M. (1979) *FEBS Lett.* **102,** 183–186.
8. Burstein, M. & Legmann, P. (1977) *Protides Biol. Fluids Proc. Colloq.* **25,** 407–410.
9. Nakaya, Y., Schaefer, E. J. & Brewer, H. B. (1980) *Biochem. Biophys. Res. Commun.* **95,** 1168–1172.
10. Schousboe, I. (1979) *Biochim. Biophys. Acta* **579,** 396–408.
11. Schousboe, I. (1980) *Thromb. Res.* **19,** 225–237.
12. Schwick, H. G. & Haupt, H. (1980) *Angew. Chem.* **92,** 83–95.
13. Finlayson, J. S. & Mushinski, J. F. (1967) *Biochim. Biophys. Acta* **147,** 413–420.
14. Liu, V. & Putnam, F. W. (1975) *Fed. Proc. Fed. Am. Soc. Exp. Biol.* **34,** 591 (abstr.).
15. Osborne, J. C., Lee, N. S. & Brewer, H. B. (1982) *Fed. Proc. Fed. Am. Soc. Exp. Biol.* **41,** 1021 (abstr.).
16. Schultze, H. E. & Heremans, J. F. (1966) *Molecular Biology of Human Proteins* (Elsevier, Amsterdam).
17. Crestfield, A. M., Moore, S. & Stein, W. H. (1963) *J. Biol. Chem.* **238,** 622–627.
18. Lozier, J. N., Takahashi, N. & Putnam, F. W. (1983) *J. Chromatogr.* **266,** 545–554.
19. Lozier, J. N. (1983) Dissertation (Indiana Univ., Bloomington, IN).
20. Dayhoff, M. O., ed. (1978) *Atlas of Protein Sequence and Structure* (National Biomedical Research Foundation, Washington, DC).
21. Barker, W. C., Hunt, L. T., Orcutt, B. C., George, D. G., Yeh, L. S., Chen, H. R., Blomquist, M. C., Johnson, G. C. & Dayhoff, M. O. (1983) *Protein Sequence Database: Atlas of Protein Sequence and Structure* (National Biomedical Research Foundation, Washington, DC), Version 7.
22. Hearn, M. T. W., Regnier, F. E. & Wehr, T. C. (1983) *High-Performance Liquid Chromatography of Proteins and Peptides* (Academic, New York).
23. Ortel, T. L., Takahashi, N. & Putnam, F. W. (1983) *J. Chromatogr.* **266,** 257–263.
24. Canfield, W. M. & Kisiel, W. (1982) *J. Clin. Invest.* **70,** 1260–1272.
25. Doolittle, R. F. (1981) *Science* **214,** 149–159.
26. Putnam, F. W., ed. (1977) *The Plasma Proteins* (Academic, New York), 2nd Ed., Vol. 3, pp. 1–284.
27. Honjo, T. (1983) *Annu. Rev. Immunol.* **1,** 499–528.
28. Lennarz, W., ed. (1980) *The Biochemistry of Glycoproteins and Proteoglycans* (Plenum, New York).