Hirabayashi, J., Kawasaki, H., Suzuki, K., & Kasai, K. (1987a) *J. Biochem. (Tokyo) 101*, 775–787.
Hirabayashi, J., Kawasaki, H., Suzuki, K., & Kasai, K. (1987b) *J. Biochem. (Tokyo) 101*, 987–995.
Huynh, T. V., Young, R. A., & Davis, R. W. (1984) *DNA Cloning Techniques: A Practical Approach* (Glover, D., Ed.) IRL Press, Oxford, U.K.
Kyte, J., & Doolittle, R. F. (1982) *J. Mol. Biol. 157*, 105–132.
Laemmli, U. K. (1970) *Nature (London) 227*, 680–685.
Maniatis, T., Fritsch, E. F., & Sambrook, J. (1982) *Molecular Cloning, A Laboratory Manual*, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY.
Merril, C. R., Goldman, F., Sedman, S. A., & Ebert, M. H. (1981) *Science (Washington, D.C.) 211*, 1437–1438.
Noelken, M. E., Nelson, C. A., Buckley, C. E., III, & Tanford, C. (1965) *J. Biol. Chem. 240*, 218–224.
Ohyama, Y., Hirabayashi, J., Oda, Y., Ohno, S., Kawasaki, H., Suzuki, K., & Kasai, K. (1986) *Biochem. Biophys. Res. Commun. 134*, 51–56.
Powell, J. T. (1980) *Biochem. J. 187*, 123–129.
Sanger, F., Nicklen, S., & Coulson, A. R. (1977) *Proc. Natl. Acad. Sci. U.S.A. 74*, 5463–5467.
Scholze, H. (1985) *J. Chromatogr. 350*, 453–460.
Southan, C., Aitken, A., Childs, R. A., Abbott, W. N., & Feizi, T. (1987) *FEBS Lett. 214*, 301–304.
Southern, E. (1975) *J. Mol. Biol. 98*, 503–517.
Sparrow, C. P., Leffler, H., & Barondes, S. H. (1987) *J. Biol. Chem. 262*, 7383–7390.
Tsunasawa, S., & Sakiyama, F. (1984) *Methods Enzymol. 106*, 165–170.
Whitney, P., Maxwell, S., Ryan, U., & Massaro, D. (1985) *Am. J. Physiol. 284*, C258–C264.
Whitney, P. L., Powell, J. T., & Sanford, G. L. (1986) *Biochem. J. 238*, 683–689.

Structural Characterization of Recombinant Hepatitis B Surface Antigen Protein by Mass Spectrometry†

Mark E. Hemling,‡ Steven A. Carr,*,‡ Carine Capiau,§ and Jean Petre§

Department of Physical and Structural Chemistry, Smith Kline & French Laboratories, P.O. Box 1539, King of Prussia, Pennsylvania 19406-0939, and Department of Human Vaccines Development, Smith Kline Biologicals, rue de l'Institute 89, B-1330 Rixensart, Belgium

Received June 18, 1987; Revised Manuscript Received September 18, 1987

ABSTRACT: The primary structure of recombinant hepatitis B surface antigen protein produced in yeast has been confirmed by mass spectrometric peptide mapping. These studies corroborate more than 85% of the amino acid sequence derived by sequencing of the gene and identified the presence of an acetyl moiety on approximately 70% of the NH_2-terminal methionine residues. Prior to the present work, direct structural analysis was largely prevented by the insolubility of this integral membrane protein and its primary degradation fragments in aqueous buffers and by partial blockage of the NH_2 terminus. These difficulties were overcome by preparative isolation using electroelution of the monomeric 226 amino acid protein from a polyacrylamide electrophoretic gel in the presence of sodium dodecyl sulfate. Chymotryptic digestion of the reduced and carboxymethylated monomer produced a large number of small, predominantly hydrophobic peptides ideally suited for peptide mapping by fast atom bombardment mass spectrometry. The percentage of NH_2-terminal methionine blocked by acetyl was determined by a new strategy involving cyanogen bromide cleavage, permethylation, and gas chromatography/mass spectrometry identification and quantitation of the *N*-methyl-*N*-acetylhomoserine produced.

Hepatitis B is a health problem of enormous magnitude afflicting more than 200 million people worldwide [for a recent review, see Tiollais et al. (1985)]. In the United States, about 0.1–2.0% of the population are carriers of the virus while in Far-East Asia and tropical Africa more than 10% of the population are carriers. Infection with hepatitis B produces a spectrum of clinical manifestations ranging from asymptomatic infection to lethal fulminant hepatitis (Junge & Deinhardt, 1985). Mortality most frequently results from chronic active hepatitis and liver disease.

The infective agent of hepatitis B belongs to a new family of hepadna viruses (Marion & Robinson, 1983). The blood of infected individuals contains three morphologically distinct carriers of the hepatitis B surface antigen (HBsAg).[1] The virion, usually present in only small amounts, is a spherical particle 42 nm in diameter consisting of an envelope and a

† This work was supported in part by l'Institut pour l'Encouragement de la Recherche dans l'Industrie et l'Agriculture (IRSIA), B-1050 Brussels, Belgium. This work was presented in part at the 34th Annual Conference on Mass Spectrometry and Allied Topics, June 8–13, 1986, Cincinnati, OH.

* Correspondence should be addressed to this author.
‡ Smith Kline & French Laboratories.
§ Smith Kline Biologicals.

[1] Abbreviations: HBsAg, hepatitis B surface antigen; HBcAg, HB core antigen; HBeAg, HB e antigen; HBV, HB virus; SDS, sodium dodecyl sulfate; PAGE, polyacrylamide gel electrophoresis; HPLC, high-performance liquid chromatography; Tris, tris(hydroxymethyl)-aminomethane; EDTA, ethylenediaminetetraacetic acid; RCM, reduced and carboxymethylated; FAB, fast atom bombardment; MS, mass spectrometry; TFA, trifluoroacetic acid; DMSO, dimethyl sulfoxide; GC, gas chromatography; PTH, phenylthiohydantoin; PTC, phenyl isothiocyanate; kDa, kilodalton(s).

0006-2960/88/0427-0699$01.50/0

nucleocapsid core. The core contains antigenic protein designated HBcAg, the viral genome, and a specific DNA polymerase. A second soluble antigenic protein designated HBeAg may also be associated with the capsid. The envelope carries the HBsAg and consists of at least three proteins present in both nonglycosylated and glycosylated forms and a lipid bilayer. The major protein, referred to as the S protein, is 226 amino acids long and is coded for by the S gene (Peterson et al., 1977). The two minor protein/glycoprotein components, pre-S2 (Stibbe & Gerlich, 1983) and pre-S1 (Heermann et al., 1984), are NH_2-terminally extended forms of the S protein containing 55 and between 163 and 174 additional amino acids, respectively. Subtypes exist which exhibit minor differences in the amino acid sequences of the proteins expressed (Tiollais et al., 1985).

The majority of HBsAg is carried by empty viral envelopes consisting of 22-nm spheres [M_r (2–4) × 10^6] and tubular structures (22 nm × 100–150 nm). These noninfectious particles represent excess surface antigen produced by the infected hepatocytes. Based primarily on amino acid analysis and SDS–PAGE, the composition of these particles is exclusively multiple copies of HBsAg proteins and glycoproteins imbedded in the lipid bilayer in roughly the same ratios as in the virions (Heermann et al., 1984).

The development of vaccines against hepatitis B virus (HBV) has focused on the immunogenic properties of the HBsAg. Until recently, investigators' attempts to propagate HBV in cell culture have failed (Sureau et al., 1986). This has led to use of vaccines based on the 22-nm particles purified from the blood of active carriers, an elaborate and expensive procedure unsuitable for mass vaccination. A more promising approach to the development of a suitable vaccine is molecular cloning and expression of the S protein coded for by the S gene of HBV DNA [for a review, see Shih (1985)]. At Smith Kline, an HBV genome (adw subtype) from the plasma of a chronic carrier was cloned and expressed in yeast (Harford et al., 1983; DeWilde et al., 1985). The host/vector system has been optimized so that the final expression plasmid contains only the minimum amount of HBV DNA necessary for expression of the S protein. Sequence analysis of the cloned S gene indicated only minor variations in sequence compared to the gene cloned by Valenzuela et al. (1982) from a virus of the same subtype. The S protein expressed is nonglycosylated and is imbedded in the lipid layer of particles similar to those found in human plasma.

Numerous difficulties have been encountered in attempts to provide direct structural confirmation of the sequence of HBsAg. Tryptic digestion of suspensions of intact antigen particles containing the reduced and aminoethylated S protein releases only three small peptides from an exposed hydrophilic domain between residues 122 and 150 (Peterson et al., 1982). Partial digestion of the protein with trypsin (with or without SDS) could only be accomplished after preparative isolation of the monomer by SDS–PAGE, exhaustive reduction, and alkylation (Peterson, 1981). However, all of the resulting fragments were insoluble (even in 6 M urea) and therefore unsuitable for analysis by automated Edman degradation. Because of these difficulties, only 23 amino acid residues in the interior of the protein have been corroborated (Peterson, 1981; Peterson et al., 1982). Furthermore, expression of the protein in yeast results in partial blockage of the NH_2 terminus which further complicates sequence analysis by Edman chemistry.

This report describes the use of nonspecific proteolytic degradation coupled with mass spectrometric peptide mapping to circumvent most of the experimental difficulties associated with structural characterization of this membrane protein [for a recent review on the use of MS in protein analysis, see Biemann and Martin (1987)]. The techniques described should be generally applicable to other highly hydrophobic proteins. These studies confirmed more than 85% of the primary structure of our recombinant HBsAg protein expressed in yeast and indicated the presence of an acetyl moiety on the NH_2 terminus of the protein. We also describe a new procedure for defining the chemical structure of blocking groups attached to NH_2-terminal Met residues and quantitating the extent of blockage. Using this strategy, we determined that approximately 70% of the protein expressed in yeast is blocked with an acetyl while the remaining 30% has a free NH_2-terminal Met.

Materials and Methods

Reagents. Chemicals and solvents used in the study were either analytical grade or HPLC grade. All electrophoresis reagents were obtained from Bio-Rad. Iodoacetic acid was a product of Janssen Pharmaceutica, Belgium. Dithiothreitol and chymotrypsin A_4 were purchased from Boehringer Mannheim GmbH, West Germany. Reagents and solvents for automated amino acid sequence analysis, all sequanal grade, were obtained from E. Merck, D-1600 Darmstadt, West Germany.

HBsAg Particles. The HBs antigen is a nonglycosylated 25-kDa protein expressed in yeast and associated in particles similar to those found in the plasma of chronic carriers. The material was produced by yeast fermentation, followed by disruption of the yeast cells containing the antigen. The HBsAg was then obtained in highly purified form by a multistep purification process involving clarification by combined precipitation and centrifugation, ultrafiltration, size-exclusion chromatography, ion-exchange chromatography, and isopycnic ultracentrifugation. The resulting HBsAg particles contained the 25-kDa polypeptide and lipids in proportions similar to plasma HBsAg (Petre et al., 1987). Their identity was established by reaction with polyclonal sera (AusRIA, Abbott Laboratories) and with monoclonal antibodies (Hauser et al., 1987). Analysis by SDS–polyacrylamide gel electrophoresis showed that they contained less than 2% yeast-derived protein contaminants.

Amino Acid Analysis. A portion (0.1 mL; 350–600 μg/mL) of each HBsAg sample was transferred into a hydrolysis tube and the solvent evaporated. The samples were extracted 1 time with 0.2 mL of chloroform/methanol/water (17:13:10) precipitating the protein. The precipitate was extracted 2 times with 0.1 mL of water and the residue dried. Aliquots of each sample were placed in each of four hydrolysis vials. Fifty microliters of 6 M HCl containing 1% phenol and 5% methyl thioglycolate was added to each vial. The vials were evacuated to about 550 millitorr and flushed with high-purity argon. This was repeated 3 times. The samples were then hydrolyzed for 24.5, 48.5, 71.5, and 94.5 h at 110 °C. At the conclusion of the hydrolysis, excess HCl was removed in vacuo, and the samples were derivatized with phenyl isothiocyanate and chromatographed on a Waters Pico-Tag system. The values were corrected for absorbance against external standards. Asp, Glu, Lys, His, Arg, Ala, Pro, Gly, and Tyr were estimated as the average of the four time points; Trp was estimated as the average from 24, 48, and 74 h; Met was measured at 48 h. Ser and Thr were estimated by extrapolation to zero time. The hydrophobic amino acids (Phe, Leu, Ile, and Val) were estimated as the average yields of the 48, 72, and 94 h of hydrolysis.

Polyacrylamide Gel Electrophoresis. Preparative polyacrylamide gel electrophoresis was performed in a Bio-Rad slab gel apparatus with 12.5% polyacrylamide gels (1.5 mm) (Laemmli, 1970). Five milligrams of purified HBsAg particles was dissolved in sample buffer, boiled for 5 min, and applied to the gel with bromophenol blue as the tracking dye. Electrophoresis was performed until the dye was eluted. Then the 25-kDa band was cut after visualization by placing the gel in 1 M cold KCl (Wallace et al., 1974). The strip was placed in a dialysis bag with 5 mL of a buffer containing 192 mM glycine, 25 mM Tris-HCl, 0.05% SDS, and 20% methanol. Electroelution took place in a Bio-Rad blotting apparatus using the same buffer. The electroelution was run twice (each time for 2 h) at 40 V (200 mA) and 4 °C. The eluted protein was dialyzed against 0.1 M ammonium bicarbonate (pH 8.2) and lyophilized.

Reduction and S-Carboxymethylation. Reduction of the HBsAg protein was performed in 0.3 M Tris-HCl buffer, pH 8.6, containing 2 mM EDTA and 8 M urea with dithiothreitol in 50-fold molar excess over cysteinyl residues. The reaction was allowed to proceed at 20 °C for 4 h under N_2 (Konigsberg, 1972). After cleavage, the HBsAg protein was immediately treated with a 100-fold molar excess of sodium iodoacetate (Hirs, 1967) at 20 °C for 20 min in the dark. The reaction was terminated by addition of β-mercaptoethanol and the mixture incubated for an additional 10 min.

N-Terminal Sequence Analysis. The polypeptide was prepared by electroelution from polyacrylamide gel slabs, reduced, and alkylated. It was treated with Extracti-Gel D suspension (Pierce) to remove SDS from the solution. After centrifugation, the sample was lyophilized. Edman degradations were performed with a Beckman 890B sequencer. Approximately 1–2 mg of protein was applied in 2% triethylamine using the standard sample application subroutine. Edman degradation was then performed with the 0.33 M Quadrol program of Hunkapillar and Hood (1978) in the presence of polybrene. The phenylthiohydantoin derivatives were quantitatively identified by high-pressure liquid chromatography on Licrosorb RP-18 (5 μM, E. Merck, Darmstadt) using a methanol gradient (20–62%, v/v) in 10 mM sodium phosphate buffer (pH 6.6).

Chymotryptic Digestion. Digestion of 2 mg (80 nmol) of the reduced and S-carboxymethylated (RCM) 25-kDa protein with chymotrypsin A_4 was performed in the volatile buffer 0.1 M ammonium bicarbonate (pH 8.2) for 20 h at 37 °C with an enzyme:substrate ratio of 1:25 (two additions of chymotrypsin at ratio of 1:50 at $t = 0$ and $t = 10$ h). The digested material was then lyophilized.

Fast Atom Bombardment Mass Spectrometry. FAB mass spectra were obtained with a VG ZAB-HF mass spectrometer equipped with a standard FAB ion source and fast atom gun. A VG 11-250 data system was used to acquire and process all data. Prior to sample analysis, a standard of bovine insulin (2 μg/μL in H_2O) was used to optimize instrument resolution and sensitivity by tuning to the peak corresponding to the A-chain fragment (2339.8 Da) and obtaining better than unit mass resolution. Approximately 1–3 μL of an enzymatic digest (approximately 1-mL total volume containing from 0.5 to 1.25 mg of digested protein) was dispersed on the stainless-steel target in a matrix of monothioglycerol (3-mercapto-1,2-propanediol). The accelerating voltage of the mass spectrometer was maintained at 8 kV while 8-keV xenon atoms at a discharge current of 1 mA were used to bombard the sample. The chymotryptic digest was analyzed in separate experiments over three different, overlapping mass ranges: (1) m/z 140–1800 at 40 s/dec; (2) m/z 900–1975 at 180 s/dec; (3) m/z 1900–3100 at 300 s/dec where dec represents decade.

Manual Edman Degradation. The digested protein (ca. 0.5 mg) was lyophilized and flushed with dry nitrogen. A 25% solution of trimethylamine (50 μL) was added through the septum and the mixture heated for 10 min at 45 °C. A 5% solution of phenyl isothiocyanate in pyridine (100 μL) was added and the mixture heated at 45 °C for 45 min. The solution was extracted twice with heptane/ethyl acetate (2:1) and dried in vacuo. Trifluoroacetic acid (100 μL) was added and the solution heated for 30 min at 40 °C. After addition of 100 μL of H_2O, the solution was extracted with 3 × 200 μL of butyl chloride and dried in vacuo.

High-Performance Liquid Chromatography. Peptide mixtures were fractionated by HPLC on a system equipped with a Beckman Model 324 pumping system (dual 114M pumps, gradient mixer, and 421 controller), a Beckman Model 210 injector, and an LDC Spectromonitor D detector, with data acquisition by a Beckman computerized automated laboratory system. Separation was done on a Waters μBondpak C18 column (10 μm, 30 cm × 3.9 mm i.d.) with a Whatman Pelicular C18 precolumn using gradient elution [1 mL/min; 100% A for 5 min and then 0–75% B over 95 min; A is 0.1% trifluoroacetic acid (TFA; sequenal grade, Pierce) in water; B is 0.07% TFA in acetonitrile (HPLC grade, Baker)]. Detection was by UV absorbance at 215 nm.

Identification of NH_2-Terminal Residue. (*A*) *Dialysis.* Monomeric HBsAg (1.76 mg, 70 nmol) in 3.9 mL of H_2O containing Tris (10 M) and SDS (0.1%) was dialyzed against 3.5 L of 0.1 M NH_4HCO_3 for 15 h at 4 °C. The sample was repetitively lyophilized to remove excess NH_4HCO_3.

(*B*) *Acetylation.* The dialyzed HBsAg was dissolved in 500 μL of [$^2H_6$]acetic anhydride and 150 μL of pyridine, purged under nitrogen, and heated for 45 min at 40 °C. Solvents were evaporated in vacuo.

(*C*) *CNBr Cleavage.* The acetylated HBsAg was dissolved in 200 μL of 70% formic acid and purged under nitrogen. Cyanogen bromide (220 μL of a 1 μmol/μL solution) was added and the solution stirred in the dark for 22 h. The reaction was quenched with 2 mL of H_2O and the sample dried in vacuo. Quenching and drying were repeated twice.

(*D*) *Permethylation.* The CNBr-cleaved HBsAg was dissolved in 50 μL of DMSO using heat (45 °C) and sonication as necessary. Methylsulfinyl carbanion (40 μL) was added to the sample under N_2 and reacted for 2 min. Methyl iodide (40 μL) was added and reacted for 10 min. Excess CH_3I was evaporated under a stream of N_2. Chloroform was added and the solution washed, first with 1 N acetic acid and then H_2O. Sample was dried in vacuo.

(*E*) *Gas Chromatography/Mass Spectrometry.* Analysis of the permethylated sample was performed by methane chemical ionization on a Finnigan 4610 GC/MS system fitted with a 15 m × 0.25 mm DB-17 capillary column. The sample (ca. 70 nmol) was dissolved in 10 μL of acetonitrile of which 2 μL was injected onto the column in a splitless mode. The spectrometer was scanned from m/z 150 to 220 at 0.5 s/scan with data acquisition by an Incos data system. Authentic [$^1H_3$]- and [$^2H_3$]acetyl-Hse standards were synthesized, and mass spectra and retention data were obtained on these materials for comparison.

RESULTS

Confirmation of Primary Structure. The amino acid sequence of the HBsAg protein deduced from sequence analysis of the S gene is shown in Figure 1. Amino acid analysis of the recombinant polypeptide (Table I) yielded compositions

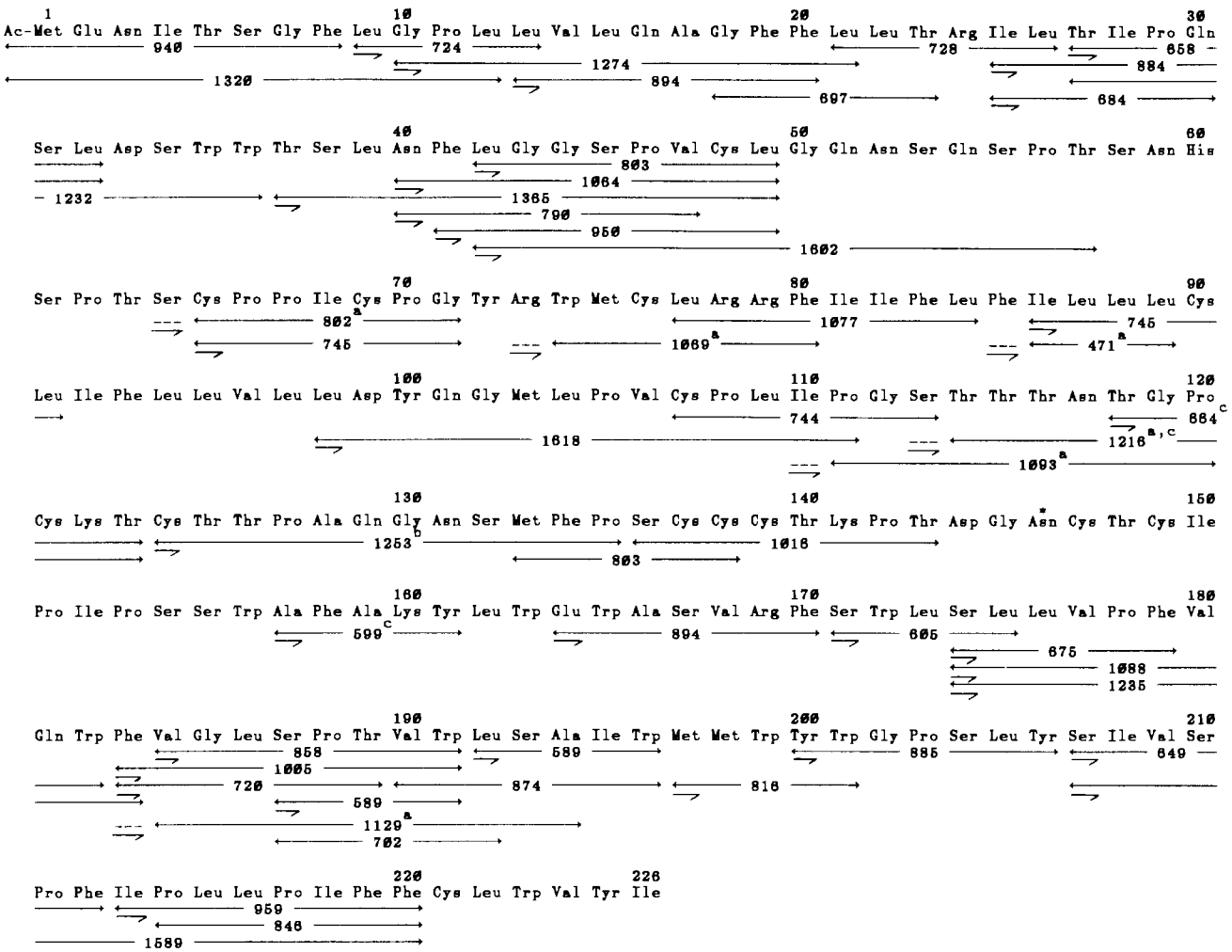

FIGURE 1: Predicted amino acid sequence of HBsAg (adw subtype) from yeast showing the peptides identified in the FABMS experiments on RCM, chymotrypsin-digested material (see text). (↔) Chymotryptic peptide; (—) amino acid released by manual Edman degradation; (superscript a) peptide signal appeared only after one step of manual Edman degradation (i.e., no precursor peptide observed) and therefore released amino acid indicated by (- - -) is inferred (see text); (superscript b) incompletely carboxymethylated peptide; (superscript c) peptides containing Lys incorporate PTC in the side chain during Edman degradation; (asterisk) potential glycosylation site. Masses correspond to the observed $(M + H)^+$.

Table I: Amino Acid Analysis of HBsAg[a]

amino acid	obsd	calcd	amino acid	obsd	calcd
Asp	9.8–10.2	10	Met	4.5–4.8	6
Thr	12.3–14.4	17	Ile	9.6–10.2[d]	16
Ser	18.9–20.5	23	Leu	23.3–23.5[d]	33
Glu	8.7–8.9	9	Tyr	5.4–5.7	6
Pro	22.4–23.1	23	Phe	10.2–10.9[d]	16
Gly	19.6–20.0[b]	14	His	1.0–1.2	1
Ala	6.4–6.8	6	Lys	2.7–3.0	3
Cys	ND[c]	14	Arg	4.6–4.9	5
Val	9.3–9.5	11	Trp	3.3–4.3[e]	13

[a] The results show the range of independent determinations on five different preparations of purified recombinant HBsAg. [b] Glycine was not accurately determined as it could not be resolved from an unidentified contaminant. [c] ND, not determined. [d] Values are consistent with previous reports (Peterson, 1981). [e] Value is low due to oxidation by incompletely reduced cystine (Gruen, 1973).

Table II: NH_2-Terminal Sequence Analysis of HBsAg[a]

step	amino acid	nmol	step	amino acid	nmol
1	Met	3.8	9	Leu	5.2
2	Glu	4.2	10	Gly	1.7
3	Asn	2.4	11	Pro	4.3
4	Ile	4.4	12	Leu	6.2
5	Thr	0.9	13	Leu	6.1
6	Ser	1.4	14	Val	5.6
7	Gly	1.6	15	Leu	5.8
8	Phe	5.6			

[a] Approximately 1–2 mg (40–80 nmol) of HBsAg polypeptide was used in the analyses. Results are expressed as nanomoles of PTH–amino acid recovered at each step.

consistent with literature values for plasma-derived HBsAg (Peterson, 1981; Gerfaux et al., 1983). Attempts to sequence the recombinant HBsAg protein monomer by automated Edman degradation gave PTH–amino acid yields ca. 5–10% for the first 15 steps (Table II), suggesting at least partial blockage of the NH_2 terminus. Substantially higher yields of PTH–amino acids (50–100%, data not shown) were obtained on the mixture of peptides resulting from CNBr cleavage of the reduced and alkylated antigen particle which is consistent with the presence of a blocking group on the NH_2-terminal Met. In contrast, NH_2-terminal analysis of plasma-derived HBsAg does not indicate any blockage and gives results typical of hydrophobic proteins, 76% for the first step with 90–92% repetitive yields for 30 steps (Peterson, 1981).

Our initial sequencing strategy was modeled after that of Peterson (1981), who demonstrated that proteolysis of HBsAg was most effective when carried out on monomeric reduced and carboxymethylated protein. Suspensions of intact antigen particles are highly resistant to proteolysis, presumably due to the hydrophobicity and degree of inter- and intramolecular

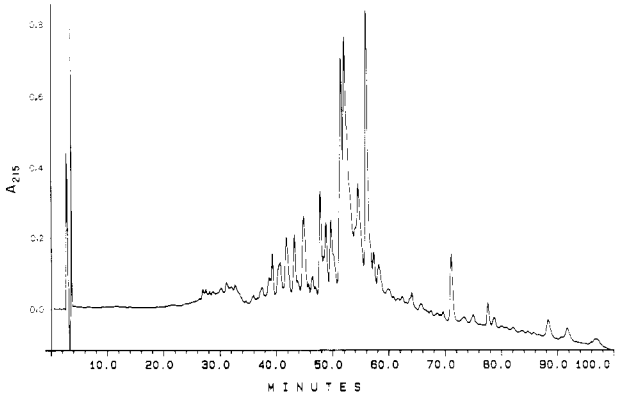

FIGURE 2: Preparative HPLC chromatogram of 10 nmol of chymotryptic digest of HBsAg. Fractions for FABMS analysis were collected at 4-min intervals from 26 min to 100 min.

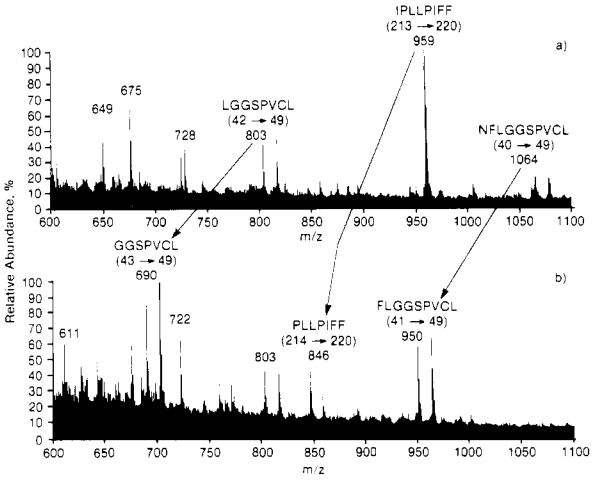

FIGURE 3: FAB mass spectra of (a) the chymotryptic digest of RCM HBsAg and (b) after one step manual Edman. Arrows indicate examples of chymotryptic and respective Edman-attenuated peptides.

disulfide bonding of the protein and the presence of lipid. Monomeric HBsAg produced by preparative SDS–PAGE followed by exhaustive reduction and carboxymethylation is still resistant to tryptic cleavage, yielding only two large insoluble peptides [M_r 12 000 and 15 000; data not shown (Peterson, 1981)]. These insoluble peptides are unsuitable for either Edman or mass spectrometric sequence analysis, and, therefore, alternative digestion strategies were investigated. The isolated polypeptide was reduced and alkylated under denaturing conditions prior to digestion with chymotrypsin. This procedure gave a large number of soluble hydrophobic peptides (Figure 2) in a mass spectrometrically useful mass range (M_r <3500). The hydrophobicity of these small peptides makes them unsuitable for automated Edman degradation as they would be readily washed from the cup within the first few steps of analysis. However, these same characteristics make them ideal candidates for analysis by the FABMS peptide mapping strategy (see below).

Analysis of a mixture of proteolytic peptides by FABMS, especially a mixture of hydrophobic peptides (Clench et al., 1985; Naylor et al., 1986), results in the observation of intense signals corresponding to the $(M + H)^+$ of peptides in the mixture. If the sequence of the digested protein is known or has been deduced from the corresponding gene, then the signals observed may be assigned to unique locations in the protein sequence based on their molecular weights and the known specificity of the enzyme employed (Biemann, 1982; Morris et al., 1983; Carr & Roberts, 1986). The complex mixture of peptides obtained by exhaustive digestion of reduced and carboxymethylated HBsAg monomer was analyzed by FABMS and mapped to the sequence of the protein by using this approach. The data from the FABMS analyses are summarized in Figure 1. Direct analysis of the chymotryptic digest provided unambiguous mapping of 53% of the sequence.

One drawback in the use of a nonspecific protease such as chymotrypsin in the FABMS peptide mapping strategy is the increased potential to produce peptides with the same nominal mass from different locations in the protein. A number of such peptide signals which could be mapped to two or more locations within the predicted sequence of the HBsAg protein were assigned to specific sequence locations after one cycle of manual Edman degradation on the mixture (Gibson & Biemann, 1984). In this procedure, the NH_2-terminal residues of each peptide in the mixture are determined by the observed changes in molecular weight of each peptide peak following one or more Edman cycles on the mixture. The mass value of each shift corresponds to the in-chain mass of the released amino acid which in turn defines where in the sequence the signal arises. For example, in Figure 3a, the peak at m/z 803 matches four peptide sequences within the protein: Ser_6–Leu_{13}, Leu_{42}–Leu_{49}, Asn_{59}–Cys_{65}, and Cys_{133}–Met_{138}. After one step of manual Edman degradation, a signal for only one of the four possible products (i.e., m/z 716, 690, 689, and 672, respectively) is observed at m/z 690 (Figure 3b) which corresponds to the loss of xLeu (Ile or Leu, in-chain mass of 113 daltons) from the m/z 803. Therefore, the peptide of $(M + H)^+ = 803$ corresponds uniquely to Leu_{42}–Leu_{49}. Likewise, the peak at m/z 959 which can also be assigned to four peptide sequences may be uniquely assigned as Ile_{213}–Phe_{220} by the presence of a peak at m/z 846 and the absence of peaks at m/z 812, 902, and 1007. The Edman mixture also produced several new peptide-related signals whose precursors were previously unobserved due to competitive suppression of their signals by other peptides and matrix components (Naylor et al., 1986). These various new assignments (peptides denoted with a superscript "a", Figure 1) increased coverage of the protein by 11% to a total of 64%.

The chymotryptic digest and the mixture subjected to Edman degradation were partially separated by HPLC to provide fractions each containing smaller numbers of peptides of similar hydrophobicity. Each of the peptide components in such a mixture is more likely to be ionized and observed (Gibson & Biemann, 1984; Naylor et al., 1986). Mass spectrometric analysis of the HPLC fractions of the proteolytic digest both before and after one cycle of manual Edman degradation confirmed many of the previously assigned peptides and provided signals for 28 previously unobserved peptides, increasing the total coverage of the protein to 85%.

Several signals observed in the various FAB mass spectra initially could not be mapped to the predicted protein sequence. Unmapped signals are important because they may represent errors in the gene sequence or its translation, the presence of significant levels of contaminating proteins, or the presence of posttranslational modifications. Two of the ion signals that could not initially be mapped to the protein were observed at m/z 940 and 1320. The largest unmapped stretch of the protein occurred at the NH_2 terminus which previous data suggested to be blocked. The difference in mass between these two peptide signals and those calculated for two possible chymotryptic fragments, Met_1–Phe_8 and Met_1–Leu_{12}, is 42 daltons, which establishes that the NH_2-terminal peptide is modified with an acetyl group and corroborates the identification made by GC/MS (vide infra). With these two signals, the data provide a contiguous sequence of 57 amino acids from the NH_2 terminus (Figure 1). Two other signals that initially

could not be mapped onto the predicted sequence of reduced and carboxymethylated HBsAg can be fit to incompletely carboxymethylated sequences (Cys_{124}–Pro_{135}, and Cys_{121}–Ala_{128} or Cys_{138}–Gly_{145}). The last ambiguous signal (m/z 894) appeared after HPLC fractionation and has been assigned as Glu_{164}–Phe_{171} in which Glu_{164} at the NH_2 terminus of the peptide has cyclized to pyroglutamic acid either during digestion or during subsequent handling. Resistance of the peptide to Edman degradation supports this assignment.

With the assignments described above, all signals observed in the FABMS data have been assigned to at least one peptide within the protein sequence contributing to the 85% coverage reported. It is generally not possible to achieve 100% coverage of a protein by FABMS using only a single proteolytic enzyme. There are several factors that limit coverage. Signals from small peptides (<600 Da) are frequently difficult to detect due to competition with ions from the liquid matrix that are particularly strong at low mass. Hydrophilic peptides respond poorly by FABMS and may be competitively suppressed by hydrophobic peptides (Clench et al., 1985; Naylor et al., 1986). Finally, the ion yield from peptides decreases with increasing mass. In the present study, either the regions of lowest coverage contained a large number of chymotryptic cleavage sites (e.g., Ile_{92}–Leu_{97}, Figure 1) and thus produced peptides too small for detection or the expected peptide was hydrophilic (e.g., Asp_{144}–Trp_{156}, Figure 1).

Chemical Identity of the NH_2-Terminal Blocking Group and the Extent of Blockage. Direct analysis of the intact protein by conventional Edman degradation required an excessive amount of material to produce sequence information, suggesting that at least some of the protein is blocked at the NH_2 terminus. The chemical nature of the blocking group and the ratio of free to blocked NH_2-terminal Met was determined by the following mass spectrometric experiments. Monomeric HBsAg was acetylated with [$^2H_6$]acetic anhydride in order to block any free NH_2 terminus with a mass spectrometrically unique derivative. The protein was then cleaved with cyanogen bromide and permethylated. The released CNBr peptides all have free NH_2 termini which quaternize during permethylation. Only the blocked (native and [$^2H_3$]-acetyl) NH_2-terminal Met (now homoserine, Hse) does not quaternize and, therefore, may be extracted into chloroform. Analysis of the chloroform extract by GC/MS revealed the presence of both [$^1H_3$]- and [$^2H_3$]acetylhomoserine as their permethyl derivatives. The mass spectra and GC retentions of these components were identical with those of authentic standards. Averaging the data over the two peaks with background subtraction indicated that ca. 70% (±10%) of the NH_2-terminal Met is blocked by an acetyl group and ca. 30% is free in the native protein derived from yeast. This procedure is specific for blocked Met and would not reveal the presence or extent of des-Met HBsAg in which Glu has been converted to pyroglutamic acid (Pca). However, analysis of the FABMS data showed no such peptides to be present.

Discussion

Establishing the sequence fidelity of recombinant proteins intended for human therapeutic use is essential to demonstrate the faithfulness of expression by the recombinant host and the absence of modifications which could result in alterations of the biological properties of the product. Unfortunately, the insolubility of membrane-associated proteins in aqueous media prevents the straightforward application of standard sequencing strategies involving proteolytic degradation and Edman-based sequence analysis. The problems encountered during attempts to sequence the particle-associated S protein are common to

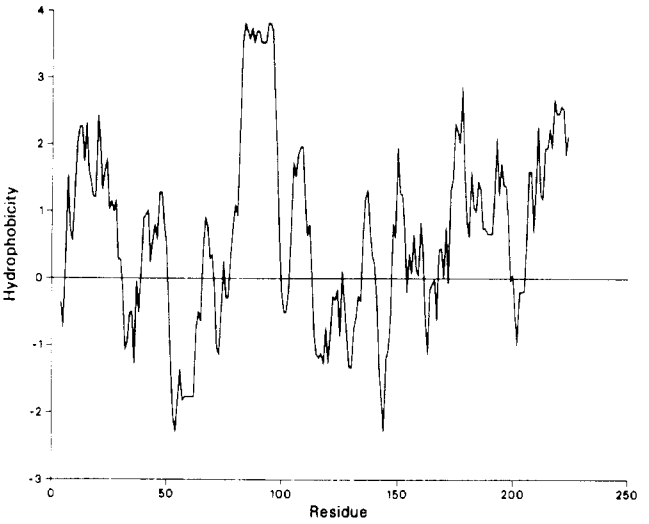

FIGURE 4: Hydropathic score (Kyte & Doolittle, 1982) of HBsAg determined by using a window of six in the PRPLOT routine of the protein identification resource.

structural studies of other integral membrane proteins (Asbeck et al., 1969; Gerber et al., 1979; Herlihy et al., 1981). The S protein (Figure 1) is exceedingly hydrophobic with an abundance of Pro, Trp, Phe, Leu, and Ile contained in three hydrophobic domains (approximately residues 6–30, 78–113, and 168–226, Figure 4), two of which likely span the lipid bilayer of the particle. Lipid constitutes more than 25% of the total mass of the particle and presumably stabilizes the protein structure through a hydrophobic interaction. The particle is further stabilized by the 14 Cys residues in each copy of the S protein all of which apparently participate in either inter- or intramolecular disulfide bonds (Peterson, 1981). Native HBsAg is not cleaved by trypsin at enzyme to protein ratios from 1:50 to 1:14; mildly reduced HBsAg particle cleaves with trypsin at only a single residue, Lys_{122}, and even following reduction and exhaustive aminoethylation of Cys residues, only a few small, soluble peptides from the apparently exposed region of the protein encompassing residues 122–147 are released (Peterson et al., 1982). Due to these difficulties, only about 25% of the protein structure had been directly established prior to this report (Peterson, 1981; Peterson et al., 1982).

Mass spectrometry is particularly useful in the analysis of membrane proteins, providing data precisely where Edman sequencing is known to fail, i.e., in the case of hydrophobic peptides and proteins. Severe "washout" of hydrophobic peptides during extraction of the PTH–amino acids greatly reduces the yields during subsequent steps in automated Edman sequencing. In the FABMS experiment, however, this hydrophobicity results in high surface activity in the liquid matrix and hence increased sensitivity toward the peptide molecular ions. While the Edman data become increasingly more difficult to read as the distance from the NH_2 terminus increases, digestion of the protein results in equally reliable FABMS data throughout the protein sequence.

In addition to hydrophobicity problems, blockage of the NH_2 terminus of HBsAg created additional difficulty in trying to obtain Edman sequence data, requiring the use of an inordinate amount of material to obtain useful information. In contrast, NH_2-terminally acetylated peptides were readily detected and identified by the FABMS peptide mapping approach. Acetylation of HBsAg is not unexpected as Met is commonly found as the terminus of NH_2-acetylated proteins (Tsunasawa & Sakiyama, 1984) and the penultimate Glu residue inhibits

cleavage of the initiator Met (Boissel et al., 1985). We believe the results presented here to be the first unambiguous demonstration of the chemical nature of a blocking group on a recombinant protein expressed in yeast. The labeling, release, and GC/MS analysis procedure developed to determine the blocking group on HBsAg is generally applicable to Met-initiated proteins and can be used to determine both the nature and the extent of blockage. We and others have demonstrated that tandem MS, which was not yet available in our laboratory during this work, is potentially a far more powerful and general approach for identifying the chemical nature of blocking groups and for sequencing posttranslationally modified proteins (Hunt et al., 1986; Crabb et al., 1986; Johnson & Biemann, 1987; Carr et al., 1988). Eventually the tandem MS technique may obviate the need for more conventional MS procedures for identifying blocking groups.

ACKNOWLEDGMENTS

We thank A. Jurewicz, B. Frederick, W. Johnson, and M. Mentzer for excellent technical assistance and Drs. G. Wasserman and P. Jeffs for many useful discussions. We also thank D. Winter and Dr. Joel Swadesh for providing the amino acid analysis.

REFERENCES

Asbeck, F., Beyreuther, K., Kohler, H., Wettstein, G., & Braunitzer, G. (1969) *Hoppe-Seyler's Z. Physiol. Chem. 350*, 1047–1066.

Biemann, K. (1982) *Int. J. Mass Spectrom. Ion Phys. 45*, 183–194.

Biemann, K., & Martin, S. A. (1987) *Mass Spectrom. Rev. 6*, 1–75.

Boissel, J.-P., Kasper, T. J., Shah, S. C., Malone, J. I., & Bunn, H. F. (1985) *Proc. Natl. Acad. Sci. U.S.A. 82*, 8448–8452.

Carr, S. A., & Roberts, G. D. (1986) *Anal. Biochem. 157*, 396–406.

Carr, S. A., Hemling, M. E., & Roberts, G. D. (1988) in *Macromolecular Sequencing and Synthesis: Selected Methods and Applications* (Schlesinger, D., Ed.) Alan R. Liss, Inc., New York (in press).

Clench, M. R., Garner, G. V., Gordon, D. B., & Barber, M. (1985) *Biomed. Mass Spectrom. 12*, 355–357.

Crabb, J. W., Armes, L. G., Carr, S. A., Johnson, C. M., Roberts, G. D., Bordoli, R. S., & McKeehan, W. L. (1986) *Biochemistry 25*, 4988–4993.

DeWilde, M., Cabezon, T., Harford, N., Rutgers, T., Simoen, E., & VanWijnendaele, F. (1985) *Dev. Biol. Stand. 59*, 99–107.

Gerber, G. E., Anderegg, R. J., Herlihy, W. C., Gray, C. P., Biemann, K., & Khorana, H. G. (1979) *Proc. Natl. Acad. Sci. U.S.A. 76*, 227–231.

Gerfaux, G., Platel, A., Adamowicz, P., & Mazert, M.-C. (1983) *Dev. Biol. Stand. 54*, 45–52.

Gibson, B. W., & Biemann, K. (1984) *Proc. Natl. Acad. Sci. U.S.A. 76*, 1956–1960.

Gruen, L. C. (1973) *Austr. J. Biol. Sci. 26*, 287–290.

Harford, N., Cabezon, T., Crabeel, M., Simoen, E., Rutgers, A., & DeWilde, M. (1983) *Dev. Biol. Stand. 54*, 125–130.

Hauser, P., Voet, P., Simoen, E., Thomas, M. C., Petre, J., DeWilde, M., & Stephenne, J. (1987) *Postgrad. Med. J. 63 (Suppl. 2)*, 83–91.

Heermann, K. H., Goldmann, U., Schwartz, W., Seyffarth, T., Baumgarten, H., & Gerlich, W. H. (1984) *J. Virol. 52*, 396–402.

Herlihy, W. C., Anderegg, R. J., & Biemann, K. (1981) *Biomed. Mass Spectrom. 8*, 62–69.

Hirs, C. H. W. (1967) *Methods Enzymol. 6*, 199–203.

Hunkapillar, M. W., & Hood, L. E. (1978) *Biochemistry 17*, 2124–2133.

Hunt, D. H., Yates, J. R., Shabanowitz, J., Winston, S., & Hauer, C. R. (1986) *Proc. Natl. Acad. Sci. U.S.A. 83*, 6233–6237.

Iwarson, S. A. (1985) in *Hepatitis B* (Gerety, R. J., Ed.) pp 119–153, Academic Press, New York.

Johnson, R. S., & Biemann, K. (1987) *Biochemistry 26*, 1209–1214.

Junge, U., & Deinhardt, F. (1985) in *Hepatitis B* (Gerety, R. J., Ed.) pp 93–118, Academic Press, New York.

Konigsberg, W. (1972) *Methods Enzymol. 25B*, 185–188.

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

Laemmli, U. K. (1970) *Nature (London) 277*, 680–681.

Marion, P. L., & Robinson, W. S. (1983) *Curr. Top. Microbiol. Immunol. 105*, 99–118.

Morris, H. R., Panico, M., & Taylor, G. W. (1983) *Biochem. Biophys. Res. Commun. 117*, 299–305.

Naylor, S., Findeis, A. F., Gibson, B. W., & Williams, D. H. (1986) *J. Am. Chem. Soc. 108*, 6359–6363.

Peterson, D. L. (1981) *J. Biol. Chem. 256*, 6975–6983.

Peterson, D. L., Roberts, I. M., & Vyas, G. N. (1977) *Proc. Natl. Acad. Sci. U.S.A. 74*, 1530–1534.

Peterson, D. L., Nath, N., & Gavilanes, F. (1982) *J. Biol. Chem. 257*, 10414–10420.

Petre, J., Van Wijnendaele, F., De Neys, R., Conrath, K., Van Opstal, O., Hauser, P., Rutgers, T., Cabezon, T., Capiau, C., Harford, N., De Wilde, M., Stephenne, J., Carr, S., Hemling, M., & Swadesh, J. (1987) *Postgrad. Med. J. 63 (Suppl. 2)*, 73–81.

Shih, J. W.-K. (1985) in *Hepatitis B* (Gerety, R. J., Ed.) pp 429–464, Academic Press, New York.

Stibbe, W., & Gerlich, W. H. (1983) *J. Virol. 46*, 626–628.

Sureau, C., Romet-Lemonne, J.-L., Millins, J. I., & Essex, H. (1986) *Cell (Cambrige, Mass.) 47*, 37–47.

Tiollais, P., Pourcel, C., & Dejean, A. (1985) *Nature (London) 317*, 489–495.

Tsunasawa, S., & Sakiayama, F. (1984) *Methods Enzymol. 106*, 165–170.

Valenzuela, P., Medina, A., Rutter, W. J., Ammerer, G., & Hall, B. D. (1982) *Nature (London) 298*, 347–350.

Wallace, R., Yu, P., Dieckert, J., & Dieckert, J. (1974) *Anal. Biochem. 61*, 86–92.