Purification and structural characterization of the putative gag-pol protease of human immunodeficiency virus.

E P Lillehoj, F H Salazar, R J Mervis, M G Raum, H W Chan, N Ahmad and S Venkatesan

***J. Virol.* 1988, 62(8):3053.**

JOURNAL OF VIROLOGY, Aug. 1988, p. 3053–3058
0022-538X/88/083053-06$02.00/0

Vol. 62, No. 8

Purification and Structural Characterization of the Putative *gag-pol* Protease of Human Immunodeficiency Virus

ERIK P. LILLEHOJ,[1] F. H. RICK SALAZAR,[2] ROBERT J. MERVIS,[3] MICHAEL G. RAUM,[4] HARDY W. CHAN,[2] NAFEES AHMAD,[3] AND SUNDARARAJAN VENKATESAN[3]*

Laboratory of Molecular Microbiology[3] and Biological Resources Branch,[4] National Institute of Allergy and Infectious Diseases, Bethesda, Maryland 20892; Program Resources, Inc., Frederick Cancer Research Facility, Frederick, Maryland 21701[1]; and Institute of Bio-Organic Chemistry, Syntex Research, Palo Alto, California 94303[2]

Received 28 January 1988/Accepted 25 April 1988

We have purified a 10,774-dalton protein from human immunodeficiency virus (HIV) type 1 that is encoded in the protease domain of the *pol* open reading frame (ORF). Radiochemical amino acid microsequencing identified 12 amino acids from the stretch of 39 N-terminal residues of this protein, beginning with a PQITLW sequence at position 69 of the *pol* ORF. Radiosequencing of selected tryptic peptides of the protein identified 11 additional residues (Leu-9 and Val-2) in six peptides encompassing the entire molecule of 99 residues. A protein of similar size and identical N-terminal sequence (determined through the first 39 residues) was present among the processed HIV *pol* gene products in *Escherichia coli* which expressed the entire HIV *pol* ORF. The C terminus of both the viral and *E. coli*-expressed proteins was inferred to be contiguous with the N terminus of the p64-p51 reverse transcriptase on the basis of tryptic mapping and specific immunoreactivity with an antiserum against a dodecapeptide located upstream of the reverse transcriptase. Thus, the initial processing of the *pol* precursor that generates the native protease is apparently preserved across phylogenetic barriers. Although the purified viral protease lacked measurable proteolytic activity, the bacterial extracts were capable of processing an HIV *gag* precursor protein synthesized in *E. coli*.

Genetic and biochemical studies have established the requirement of a specific viral protease for the expression and processing of the *gag* and *pol* gene products of several retroviruses (4, 7, 19, 20). Among the retroviruses studied so far, a viral protease is synthesized as part of the *gag* or *gag-pol* polyprotein, encoded either within the *gag* or *pol* open reading frame (ORF) or by a separate ORF overlapping both *gag* and *pol* (10–12, 15, 21). Viral proteases of 13 and 14 kilodaltons (kDa) have been purified from murine leukemia virus and bovine leukemia virus and extensively characterized (21, 22).

Many of the mature internal structural proteins of human immunodeficiency virus (HIV) type 1 are derived from the proteolytic processing of two primary translation products of 55 and of 180 to 200 kDa corresponding to the *gag* and *gag-pol* polyproteins, respectively (8, 9, 13, 16). Two forms (p64 and p51) of HIV reverse transcriptase (RT) and a p32 protein (the putative viral integration protein) juxtaposed to the C-terminal side of the p64 RT have already been sequenced (13, 17, 18). Although a candidate protein encoded by the N-terminal 167 residues of the *pol* gene that overlaps the *gag* ORF by 82 residues had not been discovered in isolated virions or infected cells, the structural homology of this domain of the *pol* ORF with other retroviral proteases prompted us to analyze extracellular virions for viral proteins using antisera against peptides corresponding to the *gag* and *pol* ORFs. By use of limited amino acid sequencing, we have identified a ca. 10-kDa protein from the virus particles and *Escherichia coli* extracts expressing the entire HIV *pol* ORF and possessing *gag* processing activity. During the course of this work, Debouck et al. have reported a *gag*-specific proteolytic activity associated with a 10-kDa protein from *E. coli* extracts which express the protease domain of the HIV genome (5).

* Corresponding author.

The *gag* and *gag-pol* gene products were analyzed by immunoprecipitation of intracellular viral proteins labeled under steady-state or pulse-chase conditions with either pooled sera from acquired immunodeficiency syndrome (AIDS) patients or rabbit hyperimmune sera raised against *E. coli* fusion proteins containing discrete structural domains of the *gag* and *pol* ORFs. Labeled extracellular viral proteins were purified by immunoprecipitation, and their partial N-terminal sequences were determined. Figure 1A illustrates the map positions of these various *gag* (S. Venkatesan, unpublished data) and *pol* (13, 18) gene products. During these early studies, a small (ca. 9 to 10 kDa) protein was consistently visualized among the immunoprecipitates with pooled sera from AIDS patients (Fig. 1B). It was occasionally immunoadsorbed by *gag* antisera and was presumed to represent a *gag* gene product. Direct N-terminal radiosequence analysis of this moiety revealed equimolar abundance of two different sequences. One was identified as beginning at position 378 of the *gag* ORF and probably represented a processed product of the C-terminal p15 *gag* protein (Fig. 1A). The other sequence determined for the 9- to 10-kDa protein was tentatively localized beginning at 69 residues from the beginning of the *pol* ORF.

The 9- to 10-kDa moiety was electroeluted from acrylamide gels and chromatographed on a DEAE-cellulose column. Successive N-terminal protein sequencing of two protein fractions eluting at 0.2 and 0.5 M NaCl was undertaken. The N-terminal sequence through 30 degradative cycles of the 0.5 M NaCl fraction exactly coincided with a stretch of residues starting at position 378 of the *gag* ORF (R. J. Mervis et al., submitted for publication). Another protein of 9 to 10 kDa was eluted with 0.2 M NaCl. To avoid confusion, the 0.2 M NaCl-eluted protein will be referred to as p10, while the *gag* protein eluting with 0.5 M NaCl will be referred to as p9. Multiple sequence analysis of the p10 protein yielded an unambiguous sequence. Figure 2a illus-

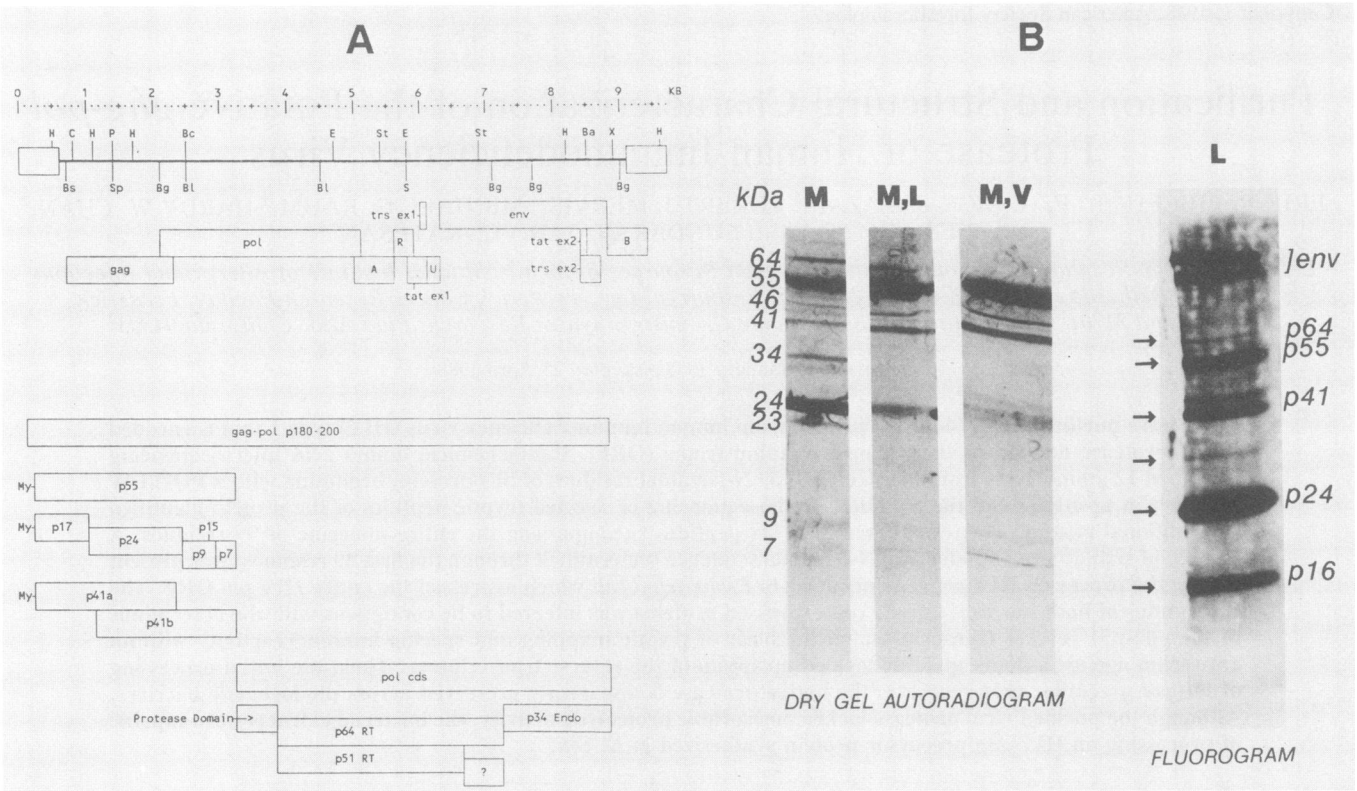

FIG. 1. Polypeptides encoded by the *gag* and *pol* genes of HIV. A recombinant HIV strain generated previously by transfection of an infectious proviral DNA, pNL432, into SW480 cells (1) was used in all the studies. For virus labeling experiments, 100 ml of A3.01 cells in RPMI 1640 medium (10^6 cells per ml) was infected with 1 ml of virus inoculum (10^5 to 10^6 50% tissure culture infective dose units) and labeled with [^{35}S]Met and the indicated ^{3}H-labeled amino acid as described previously (13). The extracellular virus was purified by precipitation with polyethylene glycol 6000 (13). In some experiments, the virus was further purified by centrifugal banding at 35,000 rpm for 90 min in a Beckman SW40 rotor at the 60% interphase of a discontinuous sucrose gradient (60, 35, and 20% [wt/vol] in 20 mM Tris hydrochloride [pH 7.5], 1 mM EDTA [TE] buffer). The virus was solubilized by the addition of NP-40 and Triton X-100 to 0.5% and dialyzed twice against 500 volumes of TE buffer with detergents before electrophoretic or chromatographic analysis. (A) Schematic illustration of the HIV genome and the relevant ORFs. The viral polypeptides derived from the *gag* and *pol* ORFs and visualized in the infected cells or purified virus are indicated in the lower part. The 180- to 200-kDa *gag-pol* precursor was identified and mapped with defined antisera in pulse-chase experiments of infected cells or transfectants expressing a series of overlapping *gag-pol* deletion plasmids (9; S. Venkatesan, unpublished data). My, myristyl group found at the N termini of p55, p41a, and p17. Mapping of the p41a and p41b *gag* proteins is based on immunoprecipitation and limited protein sequence analysis (Mervis et al., submitted). Also note that the assigned cleavage site separating p9 and p7 is tentative. The position of the protease domain is also shown. (B) Viral extracts (0.4 ml) from 2×10^8 A3.01 cells were incubated for 4 h at 4°C with 10 μl of pooled antiserum from patients with AIDS. The immune complexes were recovered by binding to protein A-Sepharose, exhaustively rinsed, and eluted for electrophoresis under reducing and denaturing conditions on 15 or 10 to 20% gradient polyacrylamide gels in SDS. Results obtained with 10 to 20% gradient acrylamide gels in SDS are shown. Lanes M; M,L; M,V; and L illustrate results obtained with virus labeled with the respective amino acid(s) (single-letter code).

trates the results obtained when p10 was labeled with Met, Leu, Lys, or Val. The experimentally determined occurrences of Leu at positions 5, 10, 19, 23, 24, 33, and 38; Lys at positions 14 and 20; Met at position 36; and Val at positions 11 and 32 precisely corresponded to their positions within a domain of the deduced sequence of the *pol* gene starting at residue 69. The probability value for the random occurrence of such a sequence was calculated (13) to be less than 1.042×10^{-19}, thus ruling out the possibility that this sequence was derived from contaminating cellular proteins.

Since direct N-terminal sequence analysis was reliable for only 35 to 40 degradative cycles, we sought to obtain the internal sequence of the p10 protein from its tryptic peptides. Individual peptides labeled with either Met and Leu or Met and Val were purified by reverse-phase high-performance liquid chromatography (RP HPLC), and their N-terminal sequences were determined. The results obtained with the Met- and Leu-labeled tryptic peptides are shown in Fig. 2b. Not all the tryptic peptides predicted by the deduced sequence of this region of the *pol* ORF were recovered by HPLC. This might have been due to the hydrophobic character of some of the internal tryptic peptides. Six Leu-labeled peptides that were ultimately recovered and sequenced are identified within the deduced sequence of the protease domain. For instance, the peptide contained in HPLC peak L2 was localized between residues 83 and 88 of the *pol* ORF, that in peak L3 was between residues 77 and 82, that in peak L7 was between residues 89 and 109, that in peak L8 was between residues 69 and 76, and that in peak L9 was between residues 114 and 123. The sequence of Val-labeled tryptic peptides further confirmed the peptide assignments shown in Fig. 2.

The HPLC-eluted viral protein had no demonstrable protease activity despite experimental variations of protein concentration (0.8 to 3.6 mg/ml), pH (2.5 to 7.0), ionic strength (0 to 0.5 M NaCl), incubation time (1 to 48 h), and temperature (22 or 37°C), conditions under which proteases of mammalian (22) and avian (2, 6, 19) retroviruses are

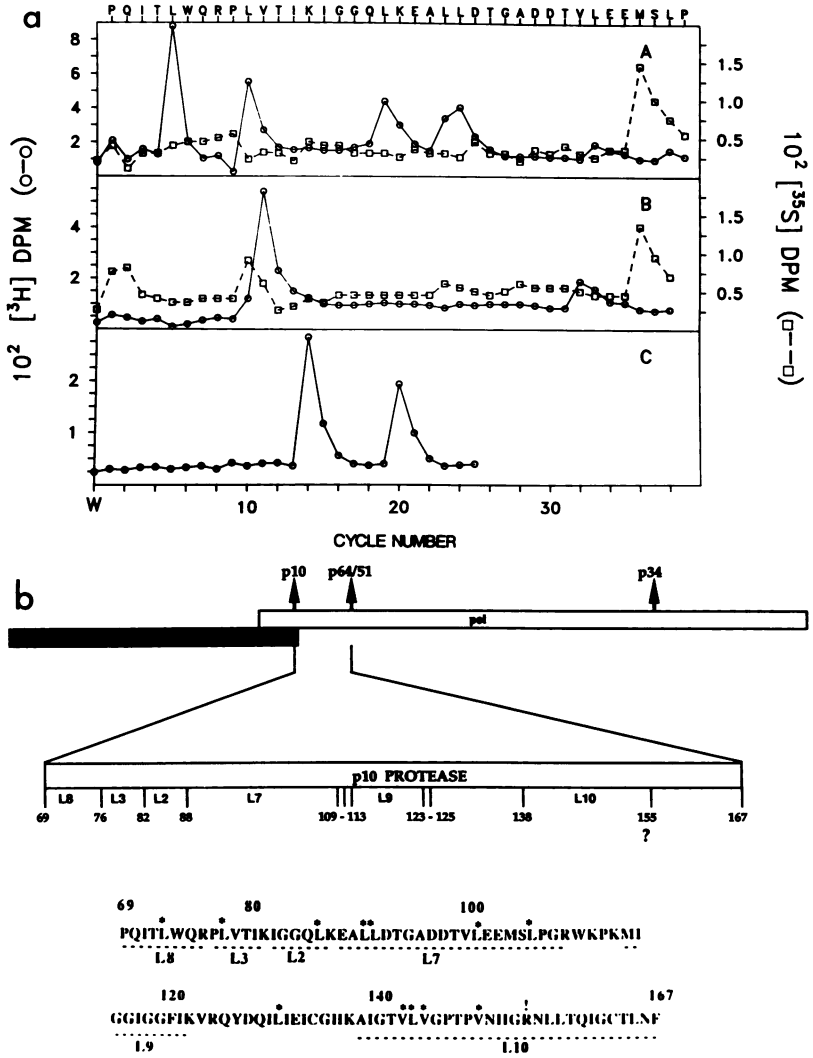

FIG. 2. N-terminal sequence analysis of the intact HIV protease and its tryptic peptides. Extracellular viral proteins eluted from the gel after immunoprecipitation were further purified by adsorption to a DEAE-cellulose column (1.0 by 5.0 cm) in TE buffer containing 0.5% NP-40 and 0.5% Triton X-100, and the column was eluted batchwise with TE buffer plus detergents containing 0.1, 0.2, 0.5, and 1.0 M NaCl. The individual fractions were analyzed by SDS-PAGE. Limited N-terminal sequencing of all fractions was undertaken. The 0.2 M NaCl fraction that yielded a homogeneous sequence of the candidate protease was mixed with 1.0 mg of crystalline bovine immunoglobulin G, reduced, alkylated, and exhaustively digested with tolylsulfonyl phenylalanyl chloromethyl ketone-treated trypsin. The tryptic digests were fractionated by RP HPLC (14) on a C_{18} column (Supelco Inc.) with a linear gradient of 0 to 100% aqueous acetonitrile containing 0.1% trifluoroacetic acid. Samples were measured for radioactivity, and selected peaks were concentrated by lyophilization and processed for N-terminal sequence determination. (a) Radiochemical sequence determination of the p10 protein (0.2 M NaCl–DEAE fraction) labeled with [^{35}S]Met and [^{3}H]Leu (A), [^{35}S]Met and [^{3}H]Val (B), or [^{3}H]Lys (C). Viral bands were visualized by wet gel autoradiography and recovered by electroelution (14). Automated N-terminal microsequence analysis by sequential Edman degradation was performed with a Beckman 890M sequenator and the Beckman 0.1 M Quadrol program 042386 (3). On the basis of the release of the individual amino acids through 39 cycles, a unique HIV protein sequence in the protease domain was identified (top line). The occurrence of [^{35}S]Met at cycles 1, 2, and 10 in profile B was attributed to a minor contaminant which constituted less than 5% of the protease-specific sequence on the basis of a repetitive yield of 96%. (b) Tryptic peptides of the putative p10 HIV protease labeled with [^{35}S]Met and either [^{3}H]Leu or [^{3}H]Val were fractionated by RP HPLC, and their partial N-terminal amino acid sequences were determined and aligned within the predicted tryptic peptides (L2, L3, L7, L8, and L10) of the protease domain. The residues identified by sequencing of the selected peptides are indicated by asterisks.

active. As an alternative, we analyzed the protease activity in extracts of *E. coli* expressing the HIV *pol* ORF, since these extracts had substantial RT activity and readily detectable amounts of mature p64-p51 and p34 HIV *pol* proteins, suggesting that the *pol* ORF gene product was processed in these cells, presumably by the HIV protease. The crude *E. coli* extracts were screened for a *gag* protease activity with a purified preparation of a truncated HIV "*gag*" precursor protein expressed in *E. coli*. This synthetic molecule was a *lacZ* fusion protein consisting of six residues of *lacZ* followed by a stretch of 348 HIV *gag*-encoded amino acids starting with residue 57 of the *gag* ORF. Cleavage of this "*gag*" substrate at the N terminus of the mature p24 *gag* protein would generate a ca. 32-kDa product containing the p24 and the N-terminal residues of the distal p9 *gag* protein. The enzymatic reactions were run for 2 or 4 h, electrophoresed on 15% polyacrylamide gels in sodium dodecyl sulfate (SDS), and analyzed by staining with Coomassie brilliant

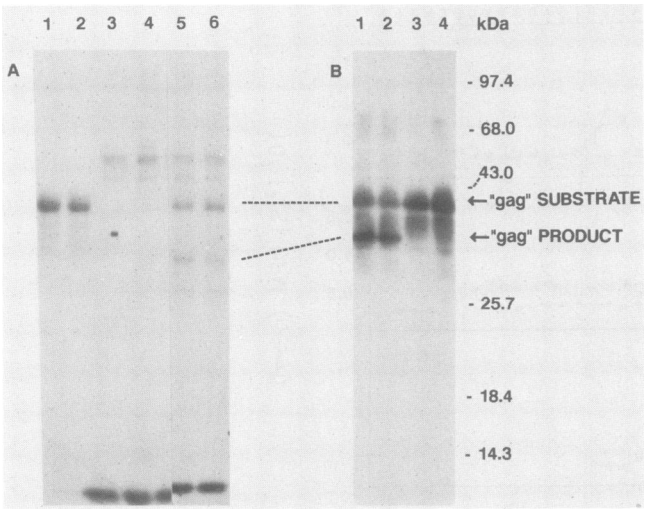

FIG. 3. Enzymatic assay of the *E. coli*-expressed protease. Bacterially expressed protease was partially purified from *E. coli* transformed by a recombinant pUC8 plasmid carrying a 3,746-base-pair *Bgl*II-*Sal*I HIV proviral DNA fragment (obtained from a subgenomic HIV proviral DNA plasmid, pBenn2 [1]) containing the entire *pol* and *sor* ORFs. The *E. coli* transformants containing the HIV *pol* ORF fused to the *lacZ* gene were induced with 1 mM isopropyl-β-D-thiogalactopyranoside for 4 h and then extracted in a buffer containing 20 mM Tris hydrochloride (pH 7.8), 1.0 M NaCl, 0.5% NP-40, and 0.5% Triton X-100. After clarification (10,000 × *g*, 10 min), the supernatants were assayed for RT activity and dialyzed exhaustively against low-salt (0.1 M NaCl) extraction buffer by using tubing with a cutoff size of 2 to 3 kDa. For assaying the HIV protease, an *E. coli* fusion protein containing a part of the HIV *gag* ORF was used. A 1,087-base-pair *Mbo*I-*Bgl*II fragment (corresponding to nucleotides 1016 to 2098 of the HIV genome) of HIV proviral DNA containing the middle one-third of the *gag* ORF was fused to the *lacZ* gene of pUC8, and protein expression was induced with isopropyl-β-D-thiogalactopyranoside. A 42-kDa *E. coli* HIV *gag* fusion protein containing the C-terminal 57 residues of p17, the entire p24, and the N-terminal 60 residues of p9 was identified by immunoblotting with *gag*-specific antisera. This protein was purified from the cell extracts to near homogeneity by two successive cycles of RP HPLC as detailed above. The 42-kDa *gag* protein was incubated at room temperature with various amounts of an extract of *E. coli* which expressed the *pol* ORF. All reactions were analyzed by SDS-PAGE and staining with Coomassie brilliant blue (A) or immunoblotting with a *gag*-specific monoclonal antibody (0058) and ^{125}I-labeled protein A (B). (A) Lanes: 1 and 2, mock assay reactions containing only 15 μg of "*gag*" substrate incubated for 2 or 4 h, respectively; 3 and 4, incubations containing only 10 μl of extract from *E. coli* expressing the *pol* ORF for 2 or 4 h, respectively; 5 and 6, complete reaction containing 10 μl of *E. coli*-expressed enzyme plus 15 μg of "*gag*" substrate incubated for 2 or 4 h, respectively. A prominent 10-kDa band seen near the bottom of the gel in *E. coli* extracts expressing the *pol* ORF probably represents the putative protease. (B) "*gag*" substrate (5 μg) was incubated for 4 h with 10 (lane 1), 2.5 (lane 2), or 0 (lane 3) μl of *E. coli* extract expressing the *pol* ORF. The substrate alone is shown in lane 4. The positions of the 42-kDa "*gag*" substrate and the 32-kDa "*gag*" reaction product are indicated by the arrows. The reactions in panels A and B were analyzed on different gels.

blue (Fig. 3A) and immunoblotting with a mouse monoclonal antibody (0058) raised against the HIV p24 *gag* protein (Fig. 3B). The crude extracts were capable of cleaving the HIV "*gag*" substrate, yielding a 32-kDa product (Fig. 3A, lanes 5 and 6, and B, lanes 1 and 2). In contrast, neither the purified substrate nor the crude extract alone produced this reaction product. Also, *E. coli* extracts expressing the vector plasmid alone had no enzymatic activity.

Partially purified bacterial extracts expressing the HIV *pol* ORF had a readily detectable band of 10 kDa (Fig. 3A, lanes 3 through 6). The extracts were fractionated by RP HPLC, and the HIV proteins in individual fractions were detected by immunoblotting with pooled sera from AIDS patients (Fig. 4A1 and 2). The bulk of the viral proteins eluting with the acetonitrile gradient included bands of 64 and 51 kDa, as expected for the HIV RT. Indeed, the fractions enriched for the p64 and p51 bands had the maximal RT activity (Fig. 4A1, fraction 30). Two additional smaller viral proteins of 10 and of 15 to 20 kDa were preferentially eluted with an isopropanol gradient (Fig. 4A2). These two proteins also specifically immunoreacted with a rabbit antiserum raised against a synthetic peptide corresponding to the C terminus of the putative viral protease (residues 154 to 167 of the HIV *pol* ORF; Fig. 4B). None of the individual HPLC fractions, however, had reproducible *gag* protease activity when the assay described above was used.

To obtain purified *E. coli*-expressed p10 protein for amino acid sequencing, the bacterial extracts labeled with [^{35}S]Met and [^{3}H]Leu or [^{3}H]Val were fractionated by RP HPLC and individual fractions were immunoprecipitated with pooled antisera from AIDS patients and resolved by SDS-polyacrylamide gel electrophoresis (PAGE) (Fig. 4C). Two proteins of 10 kDa (Fig. 4C, arrow) and of 15 to 20 kDa, eluting at ca. 40% and 55% isopropanol, respectively, immunoreacted with the pooled sera. Rabbit antisera raised against peptides corresponding to HIV *pol*-encoded residues 71 to 128 (data not shown) or 154 to 167 (Fig. 4B) also specifically reacted with these proteins. The p10 protein, labeled with Met and Leu or Met and Val, was electroeluted, and its N-terminal sequence was determined. This sequence was identical to that of the N terminus of the HIV p10 protein. Limited sequence analysis of the larger species (15 to 20 kDa) also revealed the same N terminus (data not shown). Although the C terminus of neither protein was determined, the fact that both proteins immunoreacted with rabbit antisera against peptides corresponding to the *pol*-encoded residues between 71 and 128 or 154 and 167 localized their sequences within a 99-residue domain between codons 69 and 167 of the HIV *pol* ORF. Neither of the proteins immunoreacted with monoclonal antibodies against the HIV RT sequence (data not shown).

HPLC fractions enriched for bacterially expressed p10 lacked reproducible protease activity, probably because of low recovery, poor stability, or extensive denaturation. The following lines of evidence strongly support the candidacy of p10 as the authentic HIV protease. (i) The protease activity could be detected only in *E. coli* expressing the HIV *pol* ORF and not in *E. coli* carrying the vector plasmid alone. (ii) The *gag* processing activity was abolished by antibodies directed against all or part of the protease domain (data not shown). (iii) Bacterial expression of a small DNA fragment encompassing the protease domain yielded a 10-kDa protein that processed *gag* precursor (5). (iv) *pol* ORF deletions spanning the N terminus of the protease or eliminating the entire protease domain abolished the appearance of the p64-p51 HIV RT concomitant with the disappearance of the 10-kDa protein in *E. coli* (R. Swanstrom, personal communication).

The relatively low yield of the p10 protein from virus precluded C terminus analysis by carboxypeptidase degradation. Unlabeled, HPLC-purified, *E. coli*-expressed p10 was not amenable to digestion, either. The deduced sequence of the *pol* ORF would predict a 12-residue tryptic fragment between residues 156 and 167 if the p10 and RT

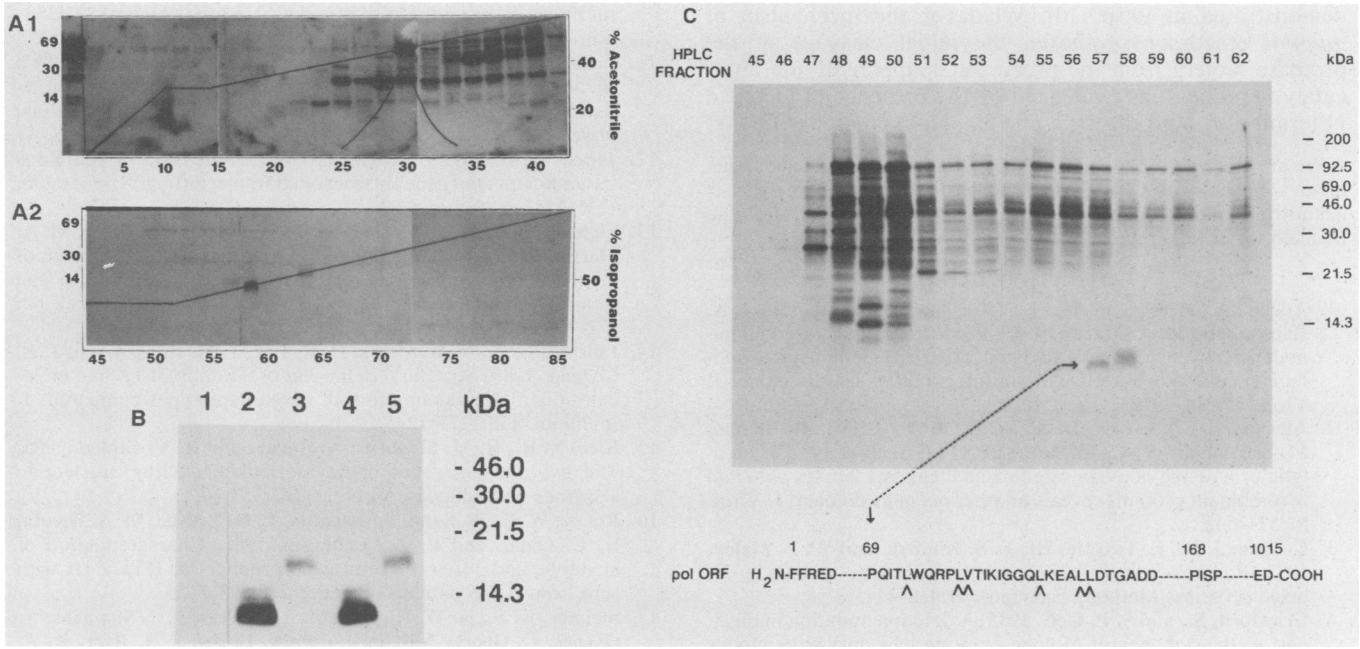

FIG. 4. Purification and amino acid sequence determination of *E. coli*-expressed protease. The dialysate of HIV *pol* ORF-expressing *E. coli* extracts (as described in the legend to Fig. 3) was applied to a 20-ml column of DEAE-cellulose in 0.1 M NaCl, and the proteins were eluted by batchwise salt gradients of 0.2, 0.5, and 1.0 M NaCl. After RT analysis and immunoblotting, the p10-enriched 0.1 to 0.2 M NaCl fractions were concentrated by acetone precipitation, dissolved in 6 M guanidine hydrochloride containing 0.1% trifluoroacetic acid (TFA), and resolved by RP HPLC on a Vydac C_4 column developed with a linear gradient of 0 to 30% aqueous acetonitrile–0.1% TFA for 30 min followed by isocratic elution for 10 min and a 60-min gradient of 30 to 60% acetonitrile–0.1% TFA. This was followed by isocratic elution for 10 min with 20% aqueous isopropanol–0.1% TFA and a 20 to 100% linear gradient of isopropanol–0.1% TFA for 60 min. Fractions (2 ml each) were collected, and samples of selected fractions were electrophoresed and immunoblotted with pooled sera from AIDS patients. (Panels A1 and A2) SDS-PAGE profiles of *E. coli* extracts which express the HIV *pol* ORF resolved by RP HPLC. The discontinuous linear acetonitrile (A1) and isopropanol (A2) gradients are denoted by lines drawn across the panels. The proteins were electroblotted and screened with pooled sera from AIDS patients and ^{125}I-labeled protein A. Selected fractions were assayed for RT activity, and the RT profile (fractions 25 to 34) is illustrated by a graph in panel A1. (B) Immunoblot detection of the two forms of the putative HIV protease expressed in *E. coli*. Two separate isopropanol gradient fractions containing the two forms of the putative HIV protease (isopropanol gradient fraction 58, lanes 2 and 4; isopropanol gradient fraction 64, lanes 3 and 5) were lyophilized and rerun on 17.5% polyacrylamide gels in SDS, electroblotted to nitrocellulose, and immunoreacted with a rabbit hyperimmune serum against a synthetic peptide containing residues 154 to 167 of the HIV *pol* ORF (lanes 2 and 3) or pooled sera from AIDS patients (lanes 4 and 5). The isopropanol fractions 58 and 64 were mixed and electrophoresed (lane 1) and reacted with nonimmune rabbit serum. (C) Immunoprecipitation of radiolabeled bacterial extracts with pooled sera from AIDS patients. A 100-ml sample of an *E. coli* culture expressing the HIV *pol* ORF was induced with 1 mM isopropyl-β-D-thiogalactopyranoside for 30 min and labeled for 2 h with a combination of [^{35}S]Met (1 mCi) and [^{3}H]Leu or [^{3}H]Val (5 mCi). Bacterial extracts were prepared and fractionated by RP HPLC as described above. Selected isopropanol gradient fractions were immunoprecipitated with pooled sera from AIDS patients and resolved by SDS-PAGE. The putative protease (arrow) was electroeluted, and its N-terminal sequence was determined (see the legend to Fig. 2); the residues identified are indicated by the carets within the relevant region of the deduced sequence of the HIV *pol* ORF below the autoradiogram. The numbers above the protein sequence refer to the residues (1 to 1012) of the HIV *pol* ORF.

were contiguous and if the protease were cleaved after Arg-155 (denoted by ! in Fig. 2b). A fragment of this size was not observed among the tryptic digestion products. However, the experimentally determined positions of Leu and Val within a large peptide, L10 (probably overlapping Arg-155), coincided with their positions in the *pol* ORF between Lys-138 and Arg-155. Among different HIV isolates and other retroviral proteases, there is a conserved domain of 6 to 9 residues surrounding an invariant Arg at position 155. We believe, therefore, that this Arg might reside within a hydrophobic pocket and be shielded from tryptic attack. Since both the viral and bacterial proteins immunoreacted with antiserum raised against a 14-residue peptide between residues 154 and 167 of the *pol* ORF, the C terminus of the protease is most likely contiguous with the N terminus of RT. On this basis the HIV protease was assumed to contain 99 amino acids with a calculated molecular mass of 10,774 and to be relatively rich in basic residues with a calculated pI of 9.83. It had a canonical -DTG- sequence (conserved among all retroviral *gag* proteases and generic aspartyl proteinases) at position 25. The protein was highly homologous to retroviral proteases over a 13-residue region centered around the -DTG- sequence. Two other lesser regions of homology with other retroviral proteases were centered around amino acid positions 52 and 87 (corresponding to the *pol* ORF residues 120 and 155, respectively).

The following three types of mechanisms have been described for the expression of the *pol* ORF of retroviruses: (i) suppression of an amber codon between the *gag* and *pol* ORFs of murine leukemia virus and feline leukemia virus (21); (ii) a single ribosomal frameshift near the start of the *pol* ORF of Rous sarcoma virus and HIV (10, 12); and (iii) double frameshifting at the junctions of the *gag-X/pro* and *X/pro-pol* ORFs of mouse mammary tumor virus (11). Since the HIV protease is contained wholly within the *pol* ORF, it is likely that frameshifting occurs either near the beginning of the *pol* ORF or immediately upstream of the protease N terminus. Frameshifting at the former site has recently been

demonstrated in vitro (10). Whatever the mechanism of *gag-pol* precursor synthesis, the initial cleavage of the protease moiety from the precursor is apparently an autocatalytic process, as evidenced by the presence of identical N termini for both the viral and *E. coli*-expressed proteins.

We are grateful to Malcolm A. Martin for support and encouragement. The help of Charles E. Buckler in computer analysis is acknowledged. Malcolm A. Martin and Arnold Rabson are also thanked for critical review of the manuscript.

LITERATURE CITED

1. **Adachi, A., H. E. Gendelman, S. Koenig, T. Folks, R. Willey, A. Rabson, and M. A. Martin.** 1986. Production of acquired immunodeficiency syndrome-associated retrovirus in human and nonhuman cells transfected with an infectious molecular clone. J. Virol. **59**:284–291.
2. **Alexander, F., J. Leis, D. A. Soltis, R. M. Crowl, W. Danho, M. S. Poonian, Y.-C. E. Pan, and A. M. Skalka.** 1987. Proteolytic processing of avian sarcoma and leukosis viruses *pol-endo* recombinant proteins reveals another *pol* gene domain. J. Virol. **61**:534–542.
3. **Coligan, J. E., F. T. Gates III, E. S. Kimball, and W. L. Maloy.** 1983. Radiochemical sequence analysis of biosynthetically labeled proteins. Methods Enzymol. **91**:413–444.
4. **Crawford, S., and S. P. Goff.** 1985. A deletion mutation in the 5′ part of the *pol* gene of Moloney murine leukemia virus blocks proteolytic processing of the *gag* and *pol* polyproteins. J. Virol. **53**:899–907.
5. **Debouck, C., J. G. Gorniak, J. E. Strickler, T. D. Meek, B. W. Metcalf, and M. Rosenberg.** 1987. Human immunodeficiency virus protease expressed in *Escherichia coli* exhibits autoprocessing and specific maturation of the *gag* precursor. Proc. Natl. Acad. Sci. USA **84**:8903–8906.
6. **Dittmar, K. J., and K. Moelling.** 1978. Biochemical properties of p15-associated protease in an avian RNA tumor virus. J. Virol. **28**:106–118.
7. **Eisenman, R. N., W. S. Mason, and M. Linial.** 1980. Synthesis and processing of polymerase proteins of wild-type and mutant avian retroviruses. J. Virol. **36**:62–78.
8. **Farmerie, W. G., D. D. Loeb, N. C. Casavant, C. A. Hutchison III, M. H. Edgell, and R. Swanstrom.** 1987. Expression and processing of the AIDS virus reverse transcriptase in *Escherichia coli*. Science **236**:305–308.
9. **Gendelman, H. E., T. S. Theodore, R. Willey, J. McCoy, A. Adachi, R. J. Mervis, S. Venkatesan, and M. A. Martin.** 1987. Molecular characterization of a polymerase mutant human immunodeficiency virus. Virology **160**:323–329.
10. **Jacks, T., M. D. Power, F. R. Masiarz, P. A. Luciw, P. J. Barr, and H. E. Varmus.** 1987. Characterization of ribosomal frameshifting in HIV-1 *gag-pol* expression. Nature (London) **331**:280–283.
11. **Jacks, T., K. Townsley, H. E. Varmus, and J. Majors.** 1987. Two efficient ribosomal frameshifting events are required for synthesis of mouse mammary tumor virus *gag*-related polyproteins. Proc. Natl. Acad. Sci. USA **84**:4298–4302.
12. **Jacks, T., and H. E. Varmus.** 1985. Expression of the Rous sarcoma virus *pol* gene by ribosomal frameshifting. Science **230**:1237–1242.
13. **Lightfoote, M. M., J. E. Coligan, T. M. Folks, A. S. Fauci, M. A. Martin, and S. Venkatesan.** 1986. Structural characterization of reverse transcriptase and endonuclease polypeptides of the acquired immunodeficiency syndrome retrovirus. J. Virol. **60**:771–775.
14. **Lillehoj, E. P., N. B. Myers, D. R. Lee, T. H. Hansen, and J. E. Coligan.** 1985. Structural definition of a family of L^d-like molecules distributed among four of seven haplotypes compared. J. Immunol. **135**:1271–1275.
15. **Rice, N. R., R. M. Stephens, A. Burny, and R. V. Gilden.** 1985. The *gag* and *pol* genes of bovine leukemia virus: nucleotide sequence and analysis. Virology **142**:357–377.
16. **Robey, W. G., B. Safai, S. Oroszlan, L. O. Arthur, M. A. Gonda, R. C. Gallo, and P. J. Fischinger.** 1985. Characterization of envelope and core structural gene products of HTLV-III with sera from AIDS patients. Science **228**:593–595.
17. **Steimer, K. S., K. W. Higgins, M. A. Powers, J. C. Stephans, A. Gyenes, C. George-Nascimento, P. A. Luciw, P. J. Barr, R. A. Hallewell, and R. Sanchez-Pescador.** 1986. Recombinant polypeptide from the endonuclease region of the acquired immune deficiency syndrome retrovirus polymerase (*pol*) gene detects serum antibodies in most infected individuals. J. Virol. **58**:9–16.
18. **Veronese, F. D., T. D. Copeland, A. L. DeVico, R. Rahman, S. Oroszlan, R. C. Gallo, and M. G. Sarngadharan.** 1986. Characterization of highly immunogenic p66/p51 as the reverse transcriptase of HTLV-III/LAV. Science **231**:1289–1291.
19. **Vogt, V. M., A. Wight, and R. Eisenman.** 1979. In vitro cleavage of avian retrovirus *gag* proteins by viral protease p15. Virology **98**:154–167.
20. **Witte, O. N., and D. Baltimore.** 1978. Relationship of retrovirus polyprotein cleavages to virion maturation studied with temperature-sensitive murine leukemia virus mutants. J. Virol. **26**:750–761.
21. **Yoshinaka, Y., I. Katoh, T. D. Copeland, and S. Oroszlan.** 1985. Murine leukemia virus protease is encoded by the *gag-pol* gene and is synthesized through suppression of an amber termination codon. Proc. Natl. Acad. Sci. USA **82**:1618–1622.
22. **Yoshinaka, Y., I. Katoh, T. D. Copeland, G. W. Smythers, and S. Oroszlan.** 1986. Bovine leukemia virus protease: purification, chemical analysis, and in vitro processing of *gag* precursor polyproteins. J. Virol. **57**:826–832.