Human immunodeficiency virus protease expressed in *Escherichia coli* exhibits autoprocessing and specific maturation of the gag precursor

(acquired immunodeficiency syndrome/RNA virus/antiviral therapy/expression vector)

CHRISTINE DEBOUCK^{†‡}, JOSELINA G. GORNIAC[§], JAMES E. STRICKLER[§], THOMAS D. MEEK[¶],
BRIAN W. METCALF[¶], AND MARTIN ROSENBERG[†]

[†]Department of Molecular Genetics, [§]Department of Macromolecular Sciences, and [¶]Department of Medicinal Chemistry, Smith Kline and French Laboratories, King of Prussia, PA 19406

Communicated by Stephen J. Benkovic, September 3, 1987 (received for review June 22, 1987)

ABSTRACT The mature gag and pol proteins of human immunodeficiency virus (HIV) and all retroviruses derive from large gag and gag-pol polyprotein precursors by posttranslational cleavage. A highly specific, virally encoded protease is required for this essential proteolytic processing. In this study, the HIV protease gene product was expressed in *Escherichia coli* and shown to autocatalyze its maturation from a larger precursor. In addition, this bacterially produced HIV protease specifically processed an HIV p55 gag polyprotein precursor when coexpressed in *E. coli*. This system will allow detailed structure-function analysis of the HIV protease and provides a simple assay for the development of potential therapeutic agents directed against this critical viral enzyme.

The human immunodeficiency virus (HIV), causative agent of acquired immunodeficiency syndrome and related disorders, is a member of the Retroviridae family (1, 2). Molecular characterization of the HIV genome has demonstrated that the virus exhibits the same overall gag-pol-env organization as other retroviruses (3–6). The HIV gag region is initially translated into a polyprotein precursor of ≈55 kDa that is then processed into the mature p17, p24, and p15 gag structural proteins (see Fig. 1A). Similarly, the gag-pol region is believed to be translated into a larger precursor through a translational frameshift between the overlapping gag and pol reading frames. This gag-pol precursor is posttranslationally processed as well, to yield the mature gag proteins and the products of the pol region including the reverse transcriptase and endonuclease.

In several retroviruses, the proteolytic maturation of the gag and gag-pol polyproteins has been shown to be effected, at least in part, by a highly specific protease that is virally encoded between the gag and pol regions (7–10). This protease is essential to the retroviral life cycle as indicated by the production of noninfectious, replication-deficient virions by Moloney murine leukemia virus variants mutated in the protease coding region (11). This suggests that specific inhibitors of the retroviral protease could block the maturation and infectivity of the virus and might be effective antiretroviral therapeutic agents, in particular, anti-HIV agents for the treatment of acquired immunodeficiency syndrome and related conditions.

To date, detailed characterization of the protease gene product from any retrovirus, and particularly HIV, has been hampered by the scarcity of this protein in virions or virus-infected cells (7–10). To obtain sufficient quantities of the HIV protease for biochemical and structural analyses, the

protease coding region was expressed in *Escherichia coli*. We report here characteristics of the expressed protease.

MATERIALS AND METHODS

Plasmids and Bacterial Strains. For the expression of gag and pro in *E. coli*, we used a derivative of the pAS expression plasmid (12, 13). In each construct, a translational fusion was created between the HIV gag or pro/pol reading frames and bacterial sequences on the expression vector. A 1286-base-pair (bp) Cla I–Bgl II restriction endonuclease fragment, including most of the p17, all of the p24, and half of the p15 gag coding sequences, was inserted in the pAS fusion vector for expression of p55 gag. To construct the protease expression vectors (PRO1–PRO4 plasmids), the following restriction endonuclease fragments were inserted in the pAS fusion vector: a 272-bp Dde I fragment for PRO1, a 259-bp Mae III–Dra I fragment for PRO2, a 382-bp NlaIV–Hae III fragment for PRO3, and a 516-bp Hae III fragment for PRO4. All restriction fragments were isolated from the BH10 clone of the HTLVIII isolate of HIV (14).

Plasmid PRO4-BX was derived from the PRO4 expression vector by digestion with Bcl I, treatment with DNA polymerase (Klenow) and ligation with the 8-mer oligonucleotide 5'-CCTCGAGG-3'. This treatment resulted in the insertion of four codons (encoding Pro-Ser-Arg-Asp) in the protease region between the conserved domains.

Plasmid PRO4-BS was derived from the PRO4 expression vector by digestion with Bcl I, treatment with DNA polymerase (Klenow), and ligation with the 12-mer oligonucleotide 5'-CTAGTTAACTAG-3', which introduces stop codons in all three reading frames. This treatment resulted in the interruption of the protease coding sequence between the conserved domains.

Plasmid pDPT287 is a derivative of the chloramphenicol-resistant pDPT101 plasmid (15). pDPT101 belongs to the incFII incompatibility group, which allows this plasmid to coexist with the ColEI-like pAS derivatives in bacteria.

The *E. coli* strain used for expression was AR120 (16). All DNA manipulations were carried out as described (17).

Protein Analyses. *E. coli* strain AR120 carrying the expression plasmids were induced, and total cell extracts were prepared and analyzed by NaDodSO₄/PAGE as described (18). Polyclonal antibodies were raised against the p55 gag and PRO1 products as described (18).

Purification and Sequencing of the Processed p24 Product. *E. coli* cells coexpressing the PRO4 and p55 gag constructs (see Fig. 3B, lane 5) were disrupted by sonication in 50 mM Tris-HCl, pH 7.5/1 mM EDTA/1 mM dithiothreitol/1 mM

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.

Abbreviation: HIV, human immunodeficiency virus.
[‡]To whom requests for reprints should be addressed.

phenylmethylsulfonyl fluoride. The soluble fraction was applied to a Mono Q high-performance anion-exchange column (Pharmacia) equilibrated in 10 mM Tris-HCl, pH 7.5. The putative p24 protein was found in the nonadsorbed fraction and subjected to reverse-phase HPLC [Brownlee RP-300 octyl column, 4.6×250 mm, equilibrated with 30% (vol/vol) acetonitrile in 0.05% trifluoroacetic acid at 1 ml/min]. The column was developed with a linear gradient to 60% (vol/vol) acetonitrile in 0.05% trifluoroacetic acid over 45 min. The p24 protein eluted in the range of 48–54% (vol/vol) acetonitrile and was $\approx 90\%$ pure as judged by Coomassie blue-stained NaDodSO₄/polyacrylamide gel. This sample was subjected to 10 cycles of automated Edman degradation in a Beckman 890M protein sequencer. Released phenylthiohydantoin-amino acid derivatives were analyzed by reverse-phase HPLC on a Beckman Ultrasphere ODS column (2×250 mm).

RESULTS AND DISCUSSION

Expression of the HIV Protease in *E. coli*. Although the primary structure of the HIV protease is not known, its coding region can be positioned between the p15 *gag* and reverse transcriptase genes by sequence comparison with other retroviruses (3–6, 19). As shown in Fig. 1, the protease coding information resides mostly within the *pol* reading frame; however, it is possible that translation of the HIV protease actually begins within the *gag* reading frame and then shifts to the *pol* frame to produce the remainder of the protein.

To express the HIV protease gene product in *E. coli*, four overlapping fragments, each encompassing the protease coding region, were inserted into a pAS bacterial expression vector (12, 13). As shown in Fig. 1B, all four fragments contained the two domains (I and II) that are highly conserved among all known retroviral proteases (19) as well as the region between these domains. The fragments differed

only in the amount of viral sequence information they carried upstream and downstream from the conserved domains. The PRO3 and PRO4 fragments extended further in both directions than did fragments PRO1 and PRO2 and, most notably, contained the proposed proteolytic cleavage site (indicated by the asterisk) Thr-Leu-Asn-Phe*Pro positioned downstream of domain II from which the amino terminus of reverse transcriptase derives (ref. 20 and Fig. 1B). In addition, the PRO3 and PRO4 fragments contained another potential protease cleavage site, Ser-Phe-Asn-Phe*Pro, positioned ≈ 20 codons upstream of conserved domain I. Proteolytic cleavage at these two sites would yield a 10-kDa polypeptide likely to correspond to the mature protease itself.

The protein products expressed by the four PRO constructs were examined initially by NaDodSO₄/PAGE. As shown in Fig. 2A, induction of the PRO1 and PRO2 constructs resulted in the accumulation of proteins with apparent molecular masses of 16 kDa and 17 kDa, respectively, which are the sizes expected for the PRO1 and PRO2 encoded products, respectively. In contrast, the PRO3 and PRO4 constructs did not produce proteins with the size expected for their encoded products (20 kDa and 25 kDa, respectively). Instead, both constructs gave rise to a protein of ≈ 10 kDa that was visualized by immunoblot analysis using a polyclonal antibody specific for the protease region (Fig. 2B). The 20-kDa and 25-kDa proteins expected from the PRO3 and PRO4 constructs could be observed for a brief period immediately after induction but rapidly disappeared to give rise to the 10-kDa protein. This apparent conversion from precursor to mature 10-kDa form occurred in the presence of chloramphenicol, a strong translation inhibitor, indicating that the 10-kDa protein results from the posttranslational processing of the larger precursors (data not shown).

Autoprocessing of the HIV Protease. The generation of a discrete 10-kDa protein by both PRO3 and PRO4 constructs is consistent with specific processing occurring at the two

FIG. 1. (A) Schematic representation of the overlapping *gag* and *pol* open reading frames (orf) and their translation products. The *gag* and *gag-pol* polyprotein precursors are outlined, and the positions of the HIV protease-mediated processing sites are indicated (*). PRO, protease; RT, reverse transcriptase; ENDO, endonuclease. (B) Structure of the portions of HIV expressed in the *E. coli* pAS expression vector. The asterisk indicates proposed processing sites for the HIV protease. Boxed regions I and II represent the two domains that are highly conserved among all known retroviral proteases (19). Restriction endonuclease sites: Bc, *Bcl* I; B, *Bgl* II; C, *Cla* I; D, *Dde* I; Dr, *Dra* I; H, *Hae* III; M, *Mae* III; N, *Nla* IV.

FIG. 2. Expression of the HIV protease coding region in bacteria. *E. coli* strain AR120 carrying the PRO constructs were induced for expression, and proteins were separated by NaDodSO₄/PAGE as described (18). Proteins were stained with Coomassie brilliant blue R-250 (A) or transferred to nitrocellulose (B) and subjected to immunoblot analysis (21) using a protease-specific rabbit polyclonal antibody raised against the PRO1 product. For the immunoblot analysis, 20 times more bacterial extract was loaded in lanes 3 and 4. Lanes: 1, PRO1; 2, PRO2; 3, PRO3; 4, PRO4; 5, PRO4-BX. The positions and sizes of prestained molecular mass markers (Bethesda Research Laboratories) are indicated in kDa.

postulated protease cleavage sites positioned on either side of the conserved domains. If the HIV protease product itself is responsible for this processing, then it suggests that this protease has an autocatalytic capability. In an effort to discriminate between bacterial proteolysis and HIV protease autoprocessing, it was reasoned that a small alteration positioned within the protease coding region far from the cleavage sites could inactivate the protease without affecting the susceptibility of the precursor protein to bacterial proteolysis. To test this, a derivative of the PRO4 expression vector was constructed in which four codons were inserted in the protease region between the conserved domains (PRO4-BX). As shown in Fig. 2, induction of this mutant derivative gave rise to a 25-kDa protein that is the size expected from the entire PRO4 coding region. No 10-kDa protein was observed in the extract. This result suggests that the small insertion destroyed the protease activity and that the processing observed truly results from the expression and autocatalytic processing of the HIV protease itself. To demonstrate directly that the 10-kDa product from induced PRO3 and PRO4 extracts indeed corresponds to the protease protein, this product was isolated and subjected to standard amino acid sequence analysis. The results indicated that the amino-terminal sequence precisely matched the sequence predicted from the proposed cleavage site located upstream of conserved domain I, Pro-Gln-Ile-Thr-Leu. The fact that the HIV protease is able to process its own precursor autocatalytically in bacteria suggests that a similar autoprocessing event is taking place within HIV-infected cells.

Processing of the HIV p55 gag Polyprotein. To further characterize the function of the bacterially expressed HIV protease, its ability to process the HIV p55 gag polyprotein precursor was examined. For this purpose, the majority of the *gag* coding region was inserted into the pAS expression vector used above (Fig. 1). The entire p55 *gag* expression unit was then excised from this vector and inserted into a second plasmid vector, pDPT287, which is known to be compatible in *E. coli* with pAS derivatives (15). Induction of cells carrying the *gag*-expressing vector alone gave rise to a 50-kDa protein, which is the size expected for the p55 protein encoded by this construct (Fig. 3A). Using the more

FIG. 3. (A) Accumulation of the p55 *gag*-related protein in *E. coli*. The AR120 strain carrying the pAS-p55 *gag* construct was induced, and total cell extracts were analyzed by NaDodSO₄/PAGE and Coomassie blue staining. Lanes: 1, pAS fusion vector without insert; 2, pAS-p55 *gag* vector. (B) Processing of the p55 *gag* polyprotein precursor by the HIV protease in *E. coli*. Each PRO construct (ampicillin resistant) and the compatible p55 *gag* expression vector (chloramphenicol resistant) were used to cotransform the AR120 strain. The cells were induced, and proteins were separated by NaDodSO₄/PAGE, transferred to nitrocellulose, and subjected to immunoblot analysis using a *gag*-specific rabbit polyclonal antibody raised against the p55 construct. Lanes: 1, pAS fusion vector without any PRO insert; 2, PRO1; 3, PRO2; 4, PRO3; 5, PRO4; 6, PRO4-BS; 7, PRO4-BX. The positions and sizes of prestained molecular mass markers (Bethesda Research Laboratories) are indicated in kDa.

sensitive immunoblot analysis with an antiserum specific for the *gag* region, the same 50-kDa major protein was observed, as well as some smaller, minor polypeptides that result presumably from bacterial proteolysis or internal translation initiations (Fig. 3B, lane 1). To examine the effect of the protein products from each of the PRO1-4 constructs on the 50-kDa *gag* precursor, bacteria were cotransformed with the p55-expressing vector and each of the PRO constructs separately. Both expression plasmids were induced simultaneously, and the extracts were analyzed by immunoblot with the *gag*-specific antiserum. As shown in Fig. 3B, lanes 4 and 5, coinduction of the p55-expressing vector and either the PRO3 or PRO4 construct resulted in the disappearance of the large 50-kDa precursor and in the concomitant appearance of two new proteins, one of ≈ 24 kDa and the other of ≈ 17 kDa. The 24-kDa product presumably corresponds to the mature p24 *gag* protein, whereas the 17-kDa protein probably contains most of the p17 *gag* region. Clearly, induction of both vectors within the same cell resulted in the processing of the p55 *gag* precursor into products with sizes consistent with cleavage at the known viral-processing sites. Finally, proper processing of purified 50-kDa *gag* precursor protein was observed *in vitro* upon incubation with an induced PRO4 bacterial extract that only contained the mature 10-kDa protease product (data not shown). To confirm that this effect was specific for the products expressed by the PRO3 and PRO4 constructs, the coinduction experiment was repeated with two PRO4 derivatives. One, PRO4-BX, contained the four-codon insertion that eliminated protease autoprocessing, and the other, PRO4-BS, contained a translation stop codon in the middle of the protease coding region. No p55 *gag* processing was observed with either of these vector constructs (Fig. 3B, lanes 6 and 7). In addition, induction of the PRO1 or PRO2 constructs also did not result in the processing of the p55-derived precursor protein. These results clearly demonstrate that the products of the PRO3 and PRO4 constructs are proteolytically active in both their precursor

Table 1. Amino-terminal amino acid sequence of the 24-kDa gag processed species

Predicted amino acid sequence	Sequence determined		
	Cycle no.	Residue*	Yield, pmol†
Pro	1	Pro	31
Ile	2	Ile	39
Val	3	Val	27
Gln	4	Gln	20
Asn	5	Asn	27
Ile	6	Ile	34
Gln	7	Gln	21
Gly	8	Gly	32
Gln	9	Gln	19
Met	10	Met	28

Predicted amino acid sequence is from ref. 22.

*Identified as phenylthiohydantoin-amino acid derivatives.

†The values shown are the actual amounts analyzed and represent 25% of the amount sequenced. The repetitive yield based on Ile-2 and Ile-6 is 96%.

and mature forms and can carry out the maturation of the gag polyprotein precursor in addition to autoprocessing.

To examine whether the p55 gag cleavage reactions observed in *E. coli* occur precisely at the sites used in viral infection, the processed 24-kDa gag protein was purified from induced cells, and its amino-terminal sequence was determined. As shown in Table 1, the first 10 amino acid residues precisely matched the amino-terminal sequence of the mature p24 gag protein isolated from viral particles (22). This demonstrates the authenticity of the p55 gag proteolytic cleavage carried out in bacteria by the 10-kDa recombinant HIV protease. It is noteworthy that the bacterially produced protease seems to be extremely active as relatively small amounts of the enzyme efficiently cleave high levels of the gag precursor. This apparently high activity observed in bacteria contrasts dramatically with the rather poor activity reported for the purified protease isolated from Rous sarcoma, murine leukemia, and bovine leukemia viral particles (7, 9, 10). Perhaps the purification procedures employed to obtain the enzyme from viral sources have impaired its activity, or perhaps the HIV protease is intrinsically a more active enzyme than those from other retroviruses.

CONCLUSIONS

The authenticity and high efficiency of the *E. coli* recombinant HIV protease product renders this system appropriate for detailed structure-function studies on the molecule. Both the autoproteolytic and selective gag precursor processing capability of this protein should allow a straightforward mutant analysis of the proteolytic function. Biochemical and physical characterization of the protein could lead to the discovery and design of potent inhibitors of this essential enzyme. This system provides an alternative approach to the potential discovery of therapeutic agents for the treatment of HIV infection.

We thank G. Sathé for synthesizing the 12-mer oligonucleotide, D. Taylor for providing the pDPT287 plasmid of incFII incompatibility, R. Gallo for the BH10 subclone, J. Young for critical reading of the manuscript, and K. Houseal and R. Hatton for editing the manuscript. This work was supported in part by Grant AI24845 from the National Institutes of Health.

- Barre-Sinoussi, F., Chermann, J. C., Rey, F., Nugeyre, M. T., Chamaret, S., Gruest, J., Dauguet, C., Axler-Blin, C., Vezinet-Brun, F., Rouzioux, C., Rozenbaum, W. & Montagnier, L. (1983) *Science* **220**, 868–870.
- Gallo, R. C., Salahuddin, S. Z., Popovic, M., Shearer, G. M., Kaplan, M., Haynes, B. F., Palker, T. J., Redfield, R., Oleske, J., Safai, B., White, G., Foster, P. & Markham, P. (1984) *Science* **224**, 500–502.
- Ratner, L., Haseltine, W. A., Patarca, R., Livak, K. J., Starcich, B., Josephs, S. F., Doran, E. R., Antoni Rafalski, J., Whitehorn, E. A., Baumeister, K., Ivanoff, L., Petteway, S. R., Pearson, M. L., Lautenberg, J. A., Papas, T. K., Ghayeb, J., Chang, N. T., Gallo, R. C. & Wong-Staal, F. (1985) *Nature (London)* **313**, 277–284.
- Wain-Hobson, S., Sonigo, P., Danos, O., Cole, S. & Alizon, M. (1985) *Cell* **40**, 9–17.
- Sanchez-Pescador, R., Power, M. D., Barr, P. J., Steimer, K. S., Stempien, M. M., Brown-Shimer, S. L., Gee, W. W., Renard, A., Randolph, A., Levy, J. A., Dina, D. & Luciw, P. A. (1985) *Science* **227**, 484–492.
- Muesing, M. A., Smith, D. H., Cabradilla, C. D., Benton, C. V., Lasky, L. A. & Capon, D. J. (1985) *Nature (London)* **313**, 450–458.
- Yoshinaka, Y., Katoh, I., Copeland, T. D. & Oroszlan, S. (1985) *Proc. Natl. Acad. Sci. USA* **82**, 1618–1622.
- Yoshinaka, Y., Katoh, I., Copeland, T. D. & Oroszlan, S. (1985) *J. Virol.* **55**, 870–873.
- Yoshinaka, Y., Katoh, I., Copeland, T. D., Smythers, G. & Oroszlan, S. (1986) *J. Virol.* **57**, 826–832.
- von der Helm, K. (1977) *Proc. Natl. Acad. Sci. USA* **74**, 911–915.
- Katoh, I., Yoshinaka, Y., Rein, A., Shibuya, M., Okada, T. & Oroszlan, S. (1985) *Virology* **145**, 280–292.
- Rosenberg, M., Ho, Y. & Shatzman, A. (1987) *Methods Enzymol.* **152**, in press.
- Shatzman, A. R. & Rosenberg, M. (1986) *Ann. N.Y. Acad. Sci.* **478**, 233–248.
- Shaw, G. M., Hahn, B. H., Arya, S. K., Groopman, J. E., Gallo, R. C. & Wong-Staal, F. (1984) *Science* **226**, 1165–1171.
- Taylor, D. P. & Cohen, S. N. (1979) *J. Bacteriol.* **137**, 92–104.
- Mott, J. E., Grant, R. A., Ho, Y. & Platt, T. (1985) *Proc. Natl. Acad. Sci. USA* **82**, 88–92.
- Maniatis, T., Fritsch, E. F. & Sambrook, J. (1982) *Molecular Cloning: A Laboratory Manual* (Cold Spring Harbor Laboratory, Cold Spring Harbor, NY).
- Aldovini, A., Debouck, C., Feinberg, M. B., Rosenberg, M., Arya, S. K. & Wong-Staal, F. (1986) *Proc. Natl. Acad. Sci. USA* **83**, 6672–6676.
- Yasunaga, T., Sagata, N. & Ikawa, Y. (1986) *FEBS Lett.* **199**, 145–150.
- di Marzo Veronese, F., Copeland, T. D., Vico, A. L., Rahman, R., Oroszlan, S., Gallo, R. C. & Sarngadharan, M. G. (1986) *Science* **231**, 1289–1291.
- Towbin, H., Staehelin, T. & Gordon, J. (1979) *Proc. Natl. Acad. Sci. USA* **76**, 4350–4354.
- Casey, J. M., Kim, Y., Andersen, P. R., Watson, K. F., Fox, J. L. & Devare, S. G. (1985) *J. Virol.* **55**, 417–423.