Analysis of Retroviral Protease Cleavage Sites Reveals Two Types of Cleavage Sites and the Structural Requirements of the P1 Amino Acid*

(Received for publication, February 19, 1991)

Steve C. Pettit‡‡, Janet Simsic‡‡, Daniel D. Loeb‡‡**, Lorraine Everitt‡, Clyde A. Hutchison III†, and Ronald Swanstrom‡ ‡***

From the ‡‡Department of Biochemistry and Biophysics, the ‡†Departments of Microbiology and Immunology and the ‡§Lineberger Comprehensive Cancer Center, University of North Carolina, Chapel Hill, North Carolina 27599

Retroviruses encode a protease which cleaves the viral Gag and Gag/Pol protein precursors into mature products. To understand the target sequence specificity of the viral protease, the amino acid sequences from 46 known processing sites from 10 diverse retroviruses were compared. Sequence preference was evident in positions P4 through P3' when compared to flanking sequences. Approximately 80% of all cleavage site sequences could be grouped into two classes based on the sequence composition flanking the scissile bond. The sequences at the amino-terminal cleavage site of the major capsid protein of Gag is always a member of one of the two classes while the carboxyl-terminal cleavage site is of the other class, suggesting a biological role for the two classes. Known processing site sequences proved useful in a motif searching strategy to identify processing sites in retroviral protein sequences, particularly in Gag. In all known cleavage sites, the P1 amino acid is hydrophobic and unbranched at the β-carbon. The sequence requirements of the P1 position were tested by site-directed mutagenesis of the P1 Phe codon in an HIV-1 Pol cleavage site. Mutations were tested for protease-mediated cleavage of the Pol precursor expressed in Escherichia coli.

Several retroviruses undergo proteolytic processing by the viral protease (PR) during virion assembly and maturation. The Gag and Gag/Pol polyprotein precursors are processed by the viral PR to generate the viral structural proteins (encoded in the gag gene) and enzymes (encoded within the pol gene) (reviewed in Refs. 1, 2). In all retroviruses, proteolytic processing of the Gag polyprotein generates the matrix protein, the capsid protein, and the nucleocapsid protein. Processing of the Gag/Pol precursor produces the reverse transcriptase, RNase H, and integrase. Depending on the retrovirus, the PR is encoded either in gag, pol, or in a small open reading frame (pro) located between gag and pol. Proteolytic processing of the precursors by the retroviral PR is an obligatory step in the formation of mature, infectious virions (3-8). HIV-1 mutants containing an inactive PR produce defective, noninfectious virions lacking the condensed core characteristic of mature virions (5, 9, 10).

Several lines of evidence indicate retroviral proteases are members of the aspartic proteinase family of proteases. However, the viral proteases differ from cellular aspartic proteinases in that they are smaller and function as a dimer (11-19). Substrate binding occurs within a cleft that is formed by symmetrical association of two PR monomers. Each monomer contains a flap that folds over the cleft that may contribute to substrate recognition and binding through interaction with the substrate (20-22).

The mechanism of processing site recognition in the precursor proteins by the viral protease is poorly understood. The PR is highly specific in site selection; only a few sites are cleaved in either the Gag or Gag/Pol precursor proteins. The cleavages are conservative in that only the scissile peptide bond is hydrolyzed at each cleavage site. Despite the specificity of site selection, recognition sequences for the retroviral PR are diverse in amino acid composition (1). Structural studies of the HIV-1 PR and substrate analogs suggest 6 to 7 residues of the substrate reside in the binding cleft (20, 22, 23). Studies with peptide substrates suggest the minimum substrate length for retroviral proteases is small (24, 25). The minimum substrate length for the HIV-1 PR corresponds to positions P4 through P3' flanking the scissile bond (26-28) (P1 is the first amino acid upstream of the scissile bond, P1' is the first amino acid downstream).

To understand cleavage site specificity, we analyzed 46 known retroviral PR cleavage site sequences from 10 different retroviruses to characterize amino acid preferences in processing site sequences. Amino acid preference was evident from position P4 through P3' flanking the scissile bond. Most significantly, the identity of the amino acid in the P1' position correlated with two distinctive patterns in the composition of the flanking amino acid sequence. Approximately 80% of the processing sites examined could be classified as one of the two types of sites. The type of processing site was conserved at certain locations in the retroviral genomes, suggesting a biological role for the two classes of sites. In addition, the sequence requirement of the P1 amino acid, which is always hydrophobic and unbranched at the β-carbon, was examined experimentally by mutagenesis of a Phe/Pro-processing site located at the amino terminus of the HIV-1 PR. Analysis of the effects of the mutations on processing of the Pol polyprotein was accomplished by an expression system in which Pol is processed to completion in Escherichia coli.
Retroviral Protease Processing Sites

MATERIALS AND METHODS

Searches of Retroviral Protein Sequences—The sequence analysis programs WEIGHTS and ANALYSEQ (29, 30) were used in the search for cleavage site sequences. Putative amino acid sequences of the gag-pol, pro, and poly open reading frames were obtained by translation of the genomic sequence of virus nucleotide sequences deposited in the EMBL/GenBank/DDBJ nucleotide sequences databases. The GenBank accession numbers of the retroviruses used are as follows: BLV: M10087, FIV: M16575, M16557, K03554, M14485; HIV-1 (HXB2D), K03455; HTLV-I, A00944, A00954, MPMV, M12349; MMTV, M15122, Mo-MuLV, J02255, J02256, J02257, RSV, J02342, J02021, J02343, SIVmac, M16403, Y00277. The citations for viral sequences can be found in the database sequence files.

Plasmid Constructions—The phagemid used for mutagenesis, sequencing, and expression of the HIV-1 pol gene was pART2. pART2 contains the pol gene from HIV-1 isolate HXB2D (31) inserted into the vector pLB120 (International Biotechnologies) (32, 33). Features of pART2 include the pol gene under transcriptional control of the lac promoter, and an Fab phage origin of replication to allow the production of single-stranded DNA for use in site-directed random mutagenesis and DNA sequencing (32, 34).

Design and Synthesis of Mutagenic DNA Oligonucleotides and the Production and Screening of Mutations in the HIV-1 pol Gene—To produce random mutations in the codon for the P1 amino acid of the Pol cleavage site located at the amino terminus of the PR domain upstream of reverse transcriptase, a 25-base DNA oligonucleotide was synthesized complementary to single-stranded PRT2 except for the three positions of the target codon where all four nucleotides were inserted with equal frequency. Single-stranded uracil-substituted pART2 for use in mutagenesis was prepared by passing the vector once through E. coli strain CJ236 (dis-"ang") (35). In vitro mutagenesis, production of pART2, and sequencing of pART2 were performed as previously described (32, 33, 35). Individual plasmid clones of the library were rescued as single-stranded DNA by infection with M13K07(b) helper phage (34). Mutations in the codon for the P1 amino acid were identified by DNA sequence analysis using the dyeoxy-chain termination method (36)

Expression of the Pol Gene—E. coli—E. coli strain JM101 containing either pART2 or pRT2 with mutations were grown overnight in modified M9 medium (no glucose, 0.4% casamino acids, 100 μg/ml ampicillin, 0.001% thiamine) (37). Expression of the pol gene from the lac promoter was induced by the addition of isopropyl-β-D-thiogalactopyranoside to a final concentration of 5 mM. After 90 min, the bacteria were pelleted by centrifugation, resuspended in 1/10 volume of sodium dodecyl sulfate-polyacrylamide gel electrophoresis loading buffer (0.25 M Tris-Cl, pH 6.8, 4% sodium dodecyl sulfate, 20% glycerol, 10% β-mercaptoethanol, 0.05% bromophenol blue), and lysed by heating to 100 °C for 3 min.

Phenotypic Screening and Western Blot Analysis—In the Western blot analysis of Pol processing, 10 μl of sample was subjected to sodium dodecyl sulfate-polyacrylamide gel electrophoresis (38) and transblotted at 70 V (300 mAmp) for 1 h onto nitrocellulose (Schleicher & Schuell BA 85) in buffer containing 25 mM Tris-Cl, pH 8.3, 192 mM glycine, and 20% (v/v) methanol. Blots were washed in phosphate-buffered saline pH 7.2 (PBS) twice for 5 min, blocked for 1 h in 3% gelatin in PBS, and exposed to 10 μl of anti-RT monoclonal (Du Pont) or anti-PR polyclonal antibody in 10 μl of PBS + 0.1% Tween 20 (PBS-Tw) for 2 h at room temperature. Blots were washed three times for 5 min in PBS-Tw and reacted with a 1:2000 dilution of either goat anti-rabbit or goat anti-mouse antibody conjugated with alkaline phosphatase in 10 μl of PBS-Tw for 1 h at room temperature (Promega Biotechnologies). The blots were then washed twice for 5 min in PBS-Tw, 5 min in Tris-buffered saline, and 5 min in carbonate buffer, pH 9.8 (0.1 M NaHCO₃, 1 mM MgCl₂). Color was developed using 5-bromo-4-chloro-3-indolyl phosphate-toluidine salt and P-nitro-blue tetrazolium as substrates for alkaline phosphatase according to the instructions of the manufacturer (Bio-Rad).

RESULTS

Two Sequence Families Correlate with the P1’ Amino Acid Sequence—To identify amino acid sequence determinants in the substrate that define PR specificity, a database containing 46 different processing site sequences from 10 highly divergent retroviruses was compiled (Table I). The cleavage sites were selected discriminately by choosing only sites proven by amino-terminal sequencing of the protein products. Several types of analyses were applied to the data base of sequences. First, substrate positions P10 through P10’ were analyzed to determine amino acid variability, the composition of amino acids, and the average hydropathicity of the amino acids located in each position (Table II). Less sequence variability was evident in positions P4 through P3’.

The four positions closest to the scissile bond (P2 through P2’?) were the most restricted in amino acid composition (Fig. 1A). The exception was the P3 position, which was as variable as positions farther away from the scissile bond. The amino acid composition in positions P10 through P10’ was also analyzed for hydrophaticity (39) (Fig. 1B). Amino acids in substrate positions P2 through P3’ were increased in hydropathicity relative to the flanking sequences. Hydrophaticity and sequence variability data, when taken together, suggested substrate sequences important for interaction with retroviral proteases generally extend from positions P4 through P3’. These positions correspond to the minimum substrate length of seven amino acids (P4 through P3’) required for cleavage by the HIV-1 PR (26, 27, 28).

The PR cleavage site sequences were also analyzed to ascertain if the presence of a certain amino acid at any one position predisposes the appearance of any particular amino acid at another position. This analysis was accomplished by correlating amino acids that appear frequently at specific positions between P5 and P5’ with the frequent appearance of specific amino acids at other positions. The analysis indicated most processing site sequences could be grouped into two classes based on two patterns of sequence preference that were evident from positions P4 through P3’. Differences in flanking amino acid composition in the two classes correlated with the amino acid in the P1’ position of the substrate.

Based on this analysis, we have designated the two classes of cleavage sites as type 1 and type 2 sites. Type 1 sites have Pro in the P1’ position (Table III), while type 2 sites have either Leu, Ala or Val in the P1’ position (Table IV). Differences in amino acid preference in the two types of sites were quite obvious. For example, type 1 sites frequently have Phe or Tyr in the P1 position while type 2 sites favor Leu over Phe or Tyr. In contrast to the differences in sequence preference between the two types of site, there were several sequence preferences that were apparently shared by both classes of sites. For example, amino acids with side chains branched at the β-carbon (Ile, Val) were excluded from the P1 position of all the processing site sequences (Table II).

The patterns of differences and similarities in the positions flanking the P1’ amino acid in the two types of PR cleavage sites can be summarized as follows: P4, small amino acids in both types of sites (Pro, Ser, Thr, Gly, Ala); P3, generally uncharged with Gln seen in some type 1 sites and Arg/Lys seen in some type 2 sites; P2’, aliphatic for both types of sites but with Asn occasionally seen and Val largely excluded from type 1 sites; P1, Tyr. Phe/Leu predominates in type 1 sites while Leu predominates in type 2 sites; P1’, Pro (type 1) and the amino acids Ala, Leu, and Val (type 2) define the two types of sites; P2’, aliphatic in both types (Ile is excluded from type 2 sites); P3’, aliphatic in both types (Ala is favored in type 2 sites).

While approximately 80% of processing site sequences were either type 1 or type 2 sites by the composition of the P1 amino acid, some PR cleavage site sequences contained amino acids other than Pro, Leu, Val, or Ala at the P1’ position. Nine such sites are listed in Table I. These sites as a group shared the general features of type 2 sites, although polar...
### TABLE I

Tabulation of known protease cleavage sites

The abbreviations used are: MA, matrix protein; CA, capsid protein; NC, nucleocapsid protein; RT, reverse transcriptase; IN, integrase; BLV, bovine leukemia virus; EIAV, equine infectious anemia virus; HTLV, human T-cell lymphotropics virus; MMTV, mouse mammary tumor virus; MP MV, Mason-Pfizer monkey virus; MuLV, Moloney murine leukemia virus; RSV, Rous sarcoma virus (Prague strain subgroup C); SIVmac, simian immuno-deficiency virus (macaque).

| Virus      | Protease cleavage site sequence | Refs |
|------------|---------------------------------|------|
| BLV       | MA/CA PPYPYDPAIL/PIISEGNNRR     | 53   |
|           | CA/NC APKVQGPAIL/VHTPGPKMPG     | 54   |
|           | x/PR PLSEALELCL/LS1PLARSRP      | 55   |
|           | PR/x PEEVPMPMV/VLDAPPShIG       | 55   |
| EIAV      | MA/CA KKSQEPSSEY/PIMIDGAGNR     | 56   |
|           | CA/x IGTTLQIKML/LAKALQGTGLA    | 56   |
|           | x/PR QKMLLAKAL/QTGLAPFKG        | 55   |
|           | NC/p9 AQRGPRQKQTF/PIQQKSQHKN   | 55   |
| HTLV-I    | MA/CA YVEPTAPQVL/VPMHPGAPP      | 42   |
|           | CA/NC WTPKDRTKVL/VVQPKPPPN      | 42   |
| HTLV-II   | MA/CA YVEPTTTCQF/PLICPPGAPS    | 57   |
|           | CA/NC WTPKDRTKVL/VVQPRPPPT      | 57   |
| HIV-1     | MA/CA GHNSNVFQNY/PIVQNQGQM      | 43   |
|           | CA/x GGPGRKARVL/EAEMQVNST      | 43   |
|           | x/NC SQVTSATIM/MQRMGRNQR       | 58   |
|           | RT/IN LVSAGIRKVL/FLDGIDKAD      | 59   |
| MMTV      | MA/pp21 ILTEQSDLVL/LSAEAKSVE    | 60   |
|           | pp21/p3 KRRKKDSKAF/LATDWNDDEL  | 60   |
|           | pp3/p6 HYHDDDELIL/PVRKRVKVK    | 60   |
|           | Pk/n RRKLPVPVGF/AGAMAEAREK     | 60   |
|           | n/CA REKGDLTFTP/VVFMGESDE      | 60   |
|           | CA/NC SPAVAVQGMAY/AAAMRGQKYS    | 60   |
|           | p30/PR EGFGSTSHV/WVEISDQ       | 60   |
| MP MV     | MA/pp21 TKKPKRFVPVL/LTAQTSKDPF | 62   |
|           | pp21/p12 KATPSATVLM/AVVNPKEELK | 62   |
|           | p12/CA KEEQIPKIDF/PVTEVTDEGQ   | 64   |
|           | CA/NC GPSYQQQGLAM/AAAFAQTVK     | 62   |
| MuLV      | MA/p12 RSTPPRSSLY/PALTPLSLAG   | 63   |
|           | p12/CA VADSTTSQAF/PLRAGNGQAL   | 63   |
|           | CA/NC RRHREMSSL/LATVSQKQD      | 63   |
|           | PR/RT GPMGQFLQLV/TLNIEDEHRL   | 64   |
|           | p15E/P2E KDRISVQQL/VLTVQYHQLK  | 65   |
| RSVMA/x   | x/p10 TPKTVGTSVC/HGCTAIGNC     | 66   |
|           | p10/CA ASTGPPVYAV/PVIVIRGQP    | 68   |
|           | CA/x LIDQGIAAAM/SSAIQPLMA     | 68   |
|           | x/NC MSSAIQPLML/AVVNEERDGQ    | 69   |
|           | NC/PR WPGPEPPAVS/LAMTMKDR     | 70   |
|           | RT/IN ADSQATFQAY/PLREAKDLH    | 71   |
| SIVmac    | MA/CA AFSGRGNGNY/PVQIQGNNT    | 67   |
|           | CA/p2 GGPGQARLML/AEALKEALAP   | 43   |
|           | p2/NC LAPAPIFPA/AAQKGPKPI     | 43   |
|           | NC/p1 MAKCPNHQAG/FLGLGPWKGK   | 43   |

Amino acids were common in the P3 position (not shown).

The argument for the existence of two classes of sites is strengthened by the observation that the classes of sites were, to a degree, conserved in their genomic location. The processing site located at the amino terminus of the capsid protein in Gag is always a type 1 site while the processing site at the carboxyl terminus of the capsid protein is usually a type 2 site and never a type 1 site (Table I). This conservation of processing site type and location is evident in all 10 retroviral genomes examined.
We were interested in developing computer methods to predict potential processing sites within Gag and Pol, based on the sequences flanking the 46 known processing sites. To preserve the relative importance of each amino acid at each position, the sequences from positions P4 to P3' were aligned relative to the scissile bond and a weight matrix was constructed from the number of occurrences of each amino acid at each position (29, 30). The ANALYSEP program performs searches for those sequences which exceed a specified score when compared to such a weight matrix (30). We began by using this program to identify potential processing sites in a database consisting of the Gag, Pol, and PR amino acid sequences of all 10 retroviruses. When the weight matrix was derived from all 46 PR cleavage sites, 7 of the top 10 and 12 of the top 20 scores corresponded to known processing sites (Fig. 2A). Approximately half of the known processing sites in the Gag, Pol, and PR sequences of the 10 viruses were identified before the search efficiency began to decrease. When searches were performed using a weight matrix derived from only the type 1 sites (16 sites) or only type 2 sites (21 sites), 8 of the top 10 scores corresponded to known cleavage site sequences. (Fig. 2, panels B and C). The processing sites identified in Gag for the 10 retroviral genomes are shown in Fig. 3A. Only those scores which fell within the efficient portion of the search (the region of the diagonal line in Fig. 2, B and C) are included. None of the nine known cleavage sites that were not classified as type 1 or 2 were identified. However, 25 of the 33 type 1 and 2 sites were identified, as were three sequences not representing known processing sites. The efficiency of cleavage site recognition in Pol was much less than for Gag (data not shown).

The success of the searches at identifying known processing sites with a fair degree of reliability encouraged us to search five additional retroviral genomes, in which processing sites have not been experimentally identified, for potential processing site sequences in Gag. The retroviral genomes examined were diverse consisting of three lentiviruses, a foamy retrovirus, and an oncovirus. The searches, conducted with data bases derived from either type 1 or 2 cleavage site sequences, identified processing site sequences of equal or better quality than many of the known cleavage site sequences. Many of the type 1 cleavage site sequences are located in genomic positions where protease cleavage sites

---

**Table II**

| P10  | P9   | P8   | P7   | P6   | P5   | P4   | P3   | P2   | P1   | P1'  | P2'  | P3'  | P4'  | P5'  | P6'  | P7'  | P8'  | P9'  | P10' |
|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|
| Cys  | 1    | 1    | 1    | 3    |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |
| Ser  | 2    | 5    | 6    | 3    | 3    | 2    | 7    | 2    | 1    | 2    | 3    | 2    | 1    | 1    | 5    | 2    | 3    | 3    | 3    |
| Thr  | 2    | 4    | 5    | 3    | 3    | 4    | 6    | 2    | 2    | 1    | 3    | 4    | 3    | 2    | 1    | 2    | 3    | 3    | 4    |
| Pro  | 4    | 9    | 6    | 8    | 6    | 9    | 12   | 1    | 1    | 16   | 5    | 7    | 6    | 4    | 6    | 4    | 9    | 2    | 4    |
| Ala  | 6    | 4    | 2    | 3    | 2    | 2    | 5    | 4    | 11   | 1    | 9    | 6    | 9    | 3    | 5    | 1    | 3    | 4    | 3    |
| Gly  | 6    | 4    | 2    | 7    | 5    | 1    | 3    | 2    | 1    | 3    | 1    | 2    | 3    | 2    | 4    | 9    | 4    | 9    | 2    |
| Asn  | 1    | 1    | 1    | 1    | 1    | 5    |      |      |      |      |      |      |      |      |      |      |      |      |      |
| Phe  | 2    | 2    | 1    | 1    | 5    | 2    | 1    | 1    |      |      |      |      |      |      |      |      |      |      |      |
| Pro  | 1    | 3    | 4    | 3    | 3    | 1    | 2    | 1    |      |      |      |      |      |      |      |      |      |      |      |
| Gin  | 1    | 2    | 2    | 3    | 5    | 6    | 9    |      |      | 1    | 3    | 6    | 6    | 2    | 2    | 6    | 1    | 6    | 1    |
| His  | 1    | 1    | 2    | 1    | 1    |      |      |      |      |      |      |      |      |      |      |      |      |      |      |
| Arg  | 4    | 3    | 1    | 2    | 1    | 4    | 2    | 3    |      |      |      |      |      |      |      |      |      |      |      |
| Lys  | 5    | 3    | 7    | 3    | 5    | 5    | 3    | 6    |      |      |      |      |      |      |      |      |      |      |      |
| Met  | 2    | 2    | 1    | 1    | 3    | 7    | 1    | 1    |      |      |      |      |      |      |      |      |      |      |      |
| Ile  | 2    | 2    | 2    | 3    |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |
| Leu  | 4    | 2    |      |      | 2    | 2    | 2    | 5    | 3    | 16   | 7    | 7    | 3    | 3    | 2    |      |      |      |      |
| Val  | 1    | 3    | 1    | 3    | 2    | 2    | 2    | 2    | 13   | 5    | 11   | 6    | 1    | 1    | 2    | 2    | 1    | 2    |
| Phe  | 2    | 2    |      |      | 2    | 2    | 10   | 2    |      |      |      |      |      |      |      |      |      |      |      |
| Tyr  | 2    | 1    | 1    | 2    |      |      | 7    |      |      |      |      |      |      |      |      |      |      |      |
| Trp  | 3    | 1    |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |

The sequence patterns associated with the type 1 and 2 sites located at the boundaries of the capsid protein are not different from other type 1 and 2 sites located elsewhere in the genome. One possible exception is the type 2 sites located at the carboxyl terminus of the capsid protein which frequently contain a basic amino acid in the P3 position. In type 2 sites located elsewhere in the genome, the P3 position is generally occupied by hydrophobic or uncharged polar amino acids.

**Identification of PR Cleavage Sites by Sequence Searching--**

A

![Graph A](image1)

B

![Graph B](image2)

**Fig. 1. Analysis of amino acid composition of retroviral protease processing sites.** A, the number of different amino acid found at each position from P10 through P10' in the 46 retroviral protease cleavage sites shown in Table I. B, sum of the hydropathy index (39) of the amino acids found in each position is plotted for the 46 retroviral processing sites. In this analysis, the sum of the hydropathic index is calculated for each position by multiplying the number of each amino acid by its hydropathic value before summation to give a weighted value based on the frequency of appearance.

The sequence patterns associated with the type 1 and 2 sites located at the boundaries of the capsid protein are not different from other type 1 and 2 sites located elsewhere in the genome. One possible exception is the type 2 sites located at the carboxyl terminus of the capsid protein which frequently contain a basic amino acid in the P3 position. In type 2 sites located elsewhere in the genome, the P3 position is generally occupied by hydrophobic or uncharged polar amino acids.

**Identification of PR Cleavage Sites by Sequence Searching--**
would be expected to occur (Fig. 3B). Searches for type 2 sites identified very few sequences when performed under the same stringent conditions as the type 1 sites. These results suggest known processing site sequences can be used with some success to identify likely processing sites in proteins, particularly type 1 sites.

**Test of P1 Specificity: Mutagenesis of the P1 Amino Acid in a Pol Cleavage Site—**Analysis of the amino acids at the P1 position of retroviral processing sites indicated hydrophobic amino acid were preferred, with Leu, Phe, Tyr, and Met being the most common (Table II). Curiously, while Leu appeared over 30% of the time, the similar amino acids Val and Ile did not appear at all (Table II). A consideration of side-chain structure of the excluded amino acids indicated that the amino acids in the P1 position have the common feature of being both uncharged and unbranched at the β-carbon.

To test the significance of this observation, amino acid substitutions were introduced by site-directed mutagenesis into the P1 position of a type 1 processing site (SNF/PQI) located at the amino terminus of the PR-encoding domain in the HIV-1 pol gene. This site, as well as the other sites in Pol, is cleaved by the PR when the entire pol gene is expressed in bacteria (32, 33, 40). The Phe codon for the P1 amino acid was mutagenized to encode 14 other amino acids. The mutated pol genes were expressed in E. coli under control of the lac promoter and the extent of processing determined by Western blot analysis (Fig. 4A). In this assay, the appearance of mature forms of the viral reverse transcriptase (64 and 51 kDa) is indicative of proper PR activity, while the size of the processed PR (normally 11 kDa) indicates whether processing has occurred at the mutated processing site. Three distinct phenotypes were observed depending on the amino acid substituted into the P1 position. Substitution of Leu for Phe permitted cleavage to occur at the mutated site at near wild-type levels as indicated by the intensity of the PR p11 band (Fig. 4B). An intermediate phenotype, characterized by reduced amounts of mature p11, was observed when Ser and Gly were substituted at the P1 position. The Ser substitution apparently reduces the amount of mature p11 more than the Gly substitution. The remainder of the mutations tested had a negative phenotype for processing to generate mature PR (Fig. 4B).
Associated with the intermediate and negative phenotypes for processing at the mutated site were the presence of several larger products (the most abundant being 25, 29, 17, and 15 kDa) that were detectable with anti-PR antibody by Western blot (Fig. 4B). The larger products were present with all amino acid substitutions tested except Leu but were absent when the pol gene contained a mutation (D25I) that eliminates protease activity (data not shown), suggesting they result from protease activity. Several of the larger protein products can be predicted from the organization of the pol gene. The 25-kDa species was probably formed by cleavage at the carboxyl terminus of PR, resulting in the retention of the amino-terminal polypeptide that is normally cleaved from the PR. Substitutions negative or reduced for processing at the amino terminus of PR also exhibited reduced processing at the other sites in Pol (Fig. 5). The amount of processing at the distal sites appeared to correlate with the amount of processing at the mutated site. While bands representing the mature reverse transcriptase p51 and p64 were present with all mutations tested, many extra bands were also evident. These bands were also present when the pol gene contained a mutation (D25I) that eliminated protease activity (PR-), suggesting they result from E. coli protease activity on the Pol precursor. These results show that substitutions that blocked processing at the mutated site reduced but did not eliminate protease activity.

**DISCUSSION**

To understand the sequence requirements that define specificity of the retroviral protease, we compared the sequences of cleavage sites from a number of retroviruses. Retroviral proteases recognize a heterogeneous collection of sequences as cleavage sites. Despite the heterogeneity, definite patterns of amino acid preference were found. Several schemes have been previously described to account for PR specificity (41–43). While the general features of those schemes are apparent in the patterns described here, the larger number of sites used here has allowed more detail to be added and has revealed the existence of two families of cleavage sites that are, in part, conserved in location in all retroviral genomes.

Because we selected sites from distantly related retroviruses, sequence specificity evolving within a family of closely related retroviruses would be overlooked in this classification.
while the class 3 sites share a P2’ Glu or Gln. The class 2 and 3 sites roughly correspond to the type 2 sites described in this broader sampling of processing sites.

The discernable patterns of sequence preference from positions P4 through P3’ flanking the scissile bond correlate with the size of the recognition sequences of the retroviral proteases. Several lines of evidence suggest that the size of the PR recognition sequence is small. Kotler et al. (24) have shown that peptides with five, but not three, amino acids on each side of the scissile bond are cleaved by the avian myeloblastosis virus PR, and peptides as short as seven amino acids are cleaved by the HIV-1 PR without significant loss of efficiency (26–28). Crystallographic structure determinations of the HIV-1 PR with a bound inhibitor suggest 6 or 7 residues of the substrate may reside in the binding cleft (20–22). In addition, we were able to delete virus-specific sequences to within four amino acids of the HIV-1 cleavage site upstream of the PR and still observe processing at that site (32). Thus, the amino acids between P4 and P3’ may represent all of the sequence information needed to permit appropriate cleavage at the scissile bond.

Two Types of Cleavage Sites Are Determined by the P1’ Amino Acid—Evidence supporting the concept of two distinct classes of cleavage site sequences was based primarily on two observations. First, in sites where the amino acid in P1’ is Pro (type 1 sites), we found distinct differences in the flanking amino acid sequence compared to sequences in which the P1’ position was occupied by Ala, Leu, or Val (type 2 sites) (Tables III and IV). Second, certain classes of cleavage sites are preferred at certain locations in all retroviral genomes. For example, the cleavage site sequence that generates the amino terminus of the major capsid protein is always a type 1 site. In contrast, the cleavage site that generates the carboxyl terminus of the major capsid protein is nearly always a type 2 site and never a type 1 site (Table I, Fig. 3A).

There are several possible reasons why two types of cleavage sites might exist. Because of the high conservation of the cleavage sites around the major capsid protein, it is possible that these sites are functionally different. For HIV-1 it has been shown that the Gag precursor undergoes an ordered series of cleavage events (44, 45) that correlates with the different rates of cleavage at the different processing sites (27, 46, 47). The rate of cleavage may be influenced not only by the amino acid composition of the cleavage site but also by the structural context and surface accessibility of the processing site. However, it is not possible at this time to determine the role secondary structure or accessibility may play in protease cleavage site sequence selection.

Degeneracy within PR Cleavage Sites—The patterns observed in amino acid preference were less apparent when examining the sites individually. Only about 50% of the type 1 or 2 cleavage site sequences listed in Table I had flanking amino acids from positions P4 to P3’ that were in complete agreement with the two general patterns for type 1 or 2 cleavage sites. The majority of the remainder of the sites differed by a single amino acid. A few sites even appeared to contain elements of both cleavage site families, for example the HIV-1 NC/p6 site has Phe-Leu in the P1–P1’ positions, and the MuLV p12E/p2E site has Val in P1’ and Gln in P3. With the exception of the cleavage sites flanking the major capsid protein, none of the other cleavage sites in Gag or Pol were conserved with respect to having a type 1 or 2 site at a specific location. It is possible that processing sites at other locations may not have the same structural and sequence constraints as the cleavage sites flanking the major capsid protein.

**Fig. 4.** A, map of pART-2 pol expression plasmid. Amino acid substitutions generated at the P1’ Phe codon of the processing site located at the amino terminus of the PR domain are indicated. pART2 contains the HIV-1 pol open reading frame under transcriptional control of the lac promoter. The plasmid also contains an f1 phage replication origin (fl IG). The sequence of the type 1 processing site located at the amino terminus of the PR domain is shown for positions P4 to P3’. Substitution mutations for the P1 amino acid (Phe) are also shown. B, Western blot of HIV-1 Pol expressed in E. coli: processing to protease. Stationary phase cultures of JM101 containing pART2 (32) or pART2 with a mutation in the codon of the P1 amino acid of the processing site located at the amino terminus of the PR domain were induced for expression of pol by the addition of isopropyl-β-D-thiogalactopyranoside. Cells were pelleted, lysed, subjected to sodium dodecyl sulfate-polyacrylamide gel electrophoresis, transferred onto nitrocellulose, and exposed to anti-PR polyclonal antibody (33). The arrows denote the position of properly processed HIV-1 protease (p11) and the unprocessed Pol precursor.

**Fig. 5.** Western blot of HIV-1 Pol expressed in E. coli: processing to reverse transcriptase. The Western blot was performed as described in Fig. 4 except anti-RT monoclonal antibody was used. The arrows indicate the position of RT p64 and p51. (PR-) indicates the pol gene contained a mutation (D251I) that renders the protease inactive (33).

of processing site sequences. An example of such a phenomenon may be the presence of Asn at P2 in type 1 sites, which appears to be restricted to primate lentiviruses (Table III). In this regard, Henderson et al. (43) examined the cleavage sites in a number of closely related primate lentiviruses and proposed the existence of three classes of cleavage sites. In their analysis, class 1 sites correspond to the type 1 sites described here, class 2 sites share a P4 Arg, a P1’ Phe, and a P2’ Leu,
Structural Requirements of the P1 Amino Acid—Of the 46 processing sites considered, eight different amino acids appeared in the P1 position (Table II). However, Phe, Tyr, Leu, and Met accounted for nearly 90% of the sites, suggesting these amino acids best fit the sequence requirements of the P1 amino acid. Thus, the specificity of the amino acid side chain in the P1 position is best described as large, hydrophobic, and unbranched at the β-carbon. The absence of the hydrophobic amino acids Ile and Val from the P1 position of cleavage sites suggested that the structure of the P1 amino acid side-chain at the β-carbon is an important determinant of cleavage site specificity.

There are several possible reasons why an amino acid with a branched β-carbon may be excluded from the P1 position. First, a branched amino acid may make hydrolysis of the peptide bond difficult, as is seen with some peptide bonds. An example is the peptide bond following the amino acids Val and Ile, which are particularity resistant to cleavage by acid hydrolysis (48). While it is generally agreed that aspartic proteinases hydrolyze peptide bonds by general base catalysis (49), there may be similar constraints by the upstream amino acid side-chains in both mechanisms of hydrolysis. Second, it is possible a β-branched side-chain on the P1 amino acid may not allow proper association of the substrate with the S1-binding pocket of the PR. Third, an unbranched amino acid may be required for some structural feature of the processing site. One possible structural constraint could involve the assumption of the cis conformation by prolines in the P1’ position. By analogy with retroviral PR cleavage sites, 50% of the prolines known to be in a cis conformation within proteins are preceded by the amino acids Phe, Tyr, and Leu, and only 3% are preceded by the branched β-carbon amino acids Ile, Val, and Thr (50).

Mutational analysis of the P1 position confirmed the sequence requirements of the P1 amino acid in a Pol cleavage site. Replacement with the β-branched amino acids Val, Ile, Pro, and Thr blocked processing at the mutated site and reduced the amount of processing at the other sites in Pol (Figs. 4B and 5). This is in agreement with an observation made by Kotler et al. (24) in which substitution of Ile in the P1 position of an avian myeloblastosis virus PR peptide substrate eliminated cleavage and the peptide functioned as an inhibitor. It is not known if the reduction of cleavage at the other processing sites in pol seen with some mutations is the result of inhibition of protease activity by the mutated sequence or by the retention of a peptide at the amino terminus of protease. The ability of Ser and Gly to substitute, at least in part, for Phe may be related to the fact that these amino acids are found in the P1 position at other processing sites, although never with Pro in P1’ (Table I). This result is in agreement with those of Partin et al. (51), in which a Ser substitution at the P1 position of the type 1 site at the amino terminus of PR in a native protein substrate allowed efficient processing in vitro by protease provided both in cis and in trans. However, a Ser substitution in the P1 position of the type 1 site located at the amino terminus of CA prevented cleavage at this site in vitro with both peptide and native protein substrates (51).

Of the amino acid substitutions we tested at the P1 position, only Leu permitted cleavage at or near wild-type levels. This suggests Leu may function as an equivalent replacement for Phe without a significant effect on cleavage. The examination of amino acid sequences from different HIV-1 isolates (52) also suggests Leu and Cys may be able to serve as functional equivalents for Phe in the P1 position of the processing site located at the amino terminus of PR. While Phe is found most often in the P1 position of this processing site in the different isolates, two isolates have Leu (HIVOYI and HIVHAN) and one isolate has Cys (HIVSF2) at the P1 position.

Acknowledgments—We thank Susan Stamper for the anti-PR antibody. We also thank Marianne Manchester and Drs. Andy Kaplan and John Olsen for critical comments about this manuscript.

REFERENCES

1. Oroszlan, S., and Luftig, T. B. (1990) in Current Topics in Microbiology and Immunology Retroviruses: Strategies of Replication (Swanstrom, R., and Vogt, P., eds) Vol. 157, pp. 153–185, Springer Verlag, Berlin
2. Swanstrom, R., Kaplan, A. H., and Manchester, M. (1990) Sem. Virol. 1, 175–186
3. Crawford, S., and Goff, S. P. (1985) J. Virol. 53, 899–907
4. Katoh, I., Yoshinaka, Y., Rein, A., Shibuya, M., Odaka, T., and Oroszlan, S. (1985) Virology 145, 280–292
5. Košl, N. E., Emini, E. A., Schleif, W. A., Davis, L. J., Heimbach, J. C., Dixon, R. A., Seedinck, E. M., and Sigal, I. S. (1989) Proc. Natl. Acad. Sci. U. S. A. 85, 4686–4690
6. Stewart, L., Schatz, G., and Vogt, V. M. (1990) J. Virol. 64, 5076–5092
7. Yoshinaka, Y., and Luftig, R. B. (1977) Proc. Natl. Acad. Sci. U. S. A. 74, 3446–3450
8. Yoshinaka, Y., and Luftig, R. B. (1978) J. Gen. Virol. 40, 151–160
9. Gottlinger, H. G., Sodroski, J. G., and Haseltine, W. A. (1989) Proc. Natl. Acad. Sci. U. S. A. 86, 5781–5785
10. Peng, C., Ho, B., Chang, T., and Chang, N. (1989) J. Virol. 63, 2550–2556
11. Katoh, I., Ikawa, Y., and Yoshinaka, Y. (1989) J. Virol. 63, 2226–2232
12. Lapatto, R., Blundell, T., Hemmings, A., Overington, J., Wilderspin, A., Wood, S., Merson, J. R., Whittle, P. J., Danley, D. E., Geoghegan, K. P., Havrylik, S. J., Lee, S. E., Schedl, K. G., and Hobart, P. M. (1989) Nature 342, 299–302
13. Meek, T. D., Dayton, B. D., Metcalf, B. W., Dreyer, G. B., Strickler, J. E., Gornik, J. G., Rosenberg, M., Moore, M. L., Magaard, V. W., and Debouck, C. (1989) Proc. Natl. Acad. Sci. U. S. A. 86, 1841–1845
14. Miller, M., Jaskolski, M., Rao, J. K. M., Leis, J., and Wodawer, A. (1989) Nature 337, 576–579
15. Navia, M. A., Fitzgerald, P. M., McKeever, B. M., Leu, C. T., Heimbach, J. C., Herber, W. K., Sigal, I. S., Darke, P. L., and Springer, J. P. (1989) Nature 337, 615–620
16. Pearl, L. H., and Taylor, W. R. (1987) Nature 329, 351–354
17. Toh, H., Ono, M., Saigo, K., and Miyata, T. (1985) Nature 315, 361
18. Wlodawer, A. M., Miller, M., Jaskolski, M., Sathyanarayana, B. K., Baldwin, E., Weber, I. T., Selk, L. M., Clawson, L., Schnei- der, J., and Kent, S. B. H. (1988) Science 245, 616–621
19. Yoshinaka, Y., and Luftig, R. B. (1980) J. Gen. Virol. 48, 329–400
20. Erickson, J., Neidhart, D. J., VanDrie, J., Kempf, D. J., Wang, X. C., Norbeck, D. W., Plattner, J. J., Rittenhouse, J. W., Turon, M., Wideburg, N., Kohlbrenner, W. E., Simmer, R., Eichrich, R., Paul, D. A., and Knigge, M. (1988) Science 249, 527–533
21. Fitzgerald, P. M. D., McKeever, B. M., VanMiddlesworth, J. F., Springer, J. P., Heimbach, J. C., Leu, C.-T., Herber, W. K., Dixon, R. A. F., and Darke, P. L. (1989) J. Biol. Chem. 265, 14295–14219
22. Miller, M., Schneider, J., Sathyanarayana, B. K., Toth, M. V., Marshall, G. R., Clawson, L., Selk, L., Kent, S. B. H., and Wodawer, A. (1989) Science 246, 1149–1152
23. Weber, I. T., Miller, M., Jaskolski, M., Leis, J., Skalka, A. M., and Wodawer, A. (1989) Science 245, 928–31
24. Kotler, M., Katz, R. A., Danho, W., Leis, J., and Skalka, A. M. (1988) Proc. Natl. Acad. Sci. U. S. A. 85, 4185–4189
25. Kotler, M., Danho, W., Katz, R. A., Leis, J., and Skalka, A. M. (1989) J. Biol. Chem. 264, 3428–3435
26. Billings, S., Knoop, M. T., Hansen, J., Strup, P., Sedlacek, J., Mertz, R., and Moelling, K. (1988) J. Biol. Chem. 263, 17905–17908
27. Darke, P. L., Nutt, R. F., Brady, S. F., Garsky, V. M., Cicerone, T. M., Leu, C. T., Lumma, P. K., Friedinger, R. M., Veber, D. F., and Sigal, I. (1988) *Biochem. Biophys. Res. Commun.* 156, 297-303

28. Moore, M. L., Bryan, W. M., Fakhoury, S. A., Magada, V. W., Huffman, W. F., Dayton, B. D., Meek, T. D., Hyland, L., Dreyer, G. B., Metcalf, B. W., Strickler, J. E., Gorniak, J. G., and Debouck, C. (1989) *Biochem. Biophys. Res. Commun.* 159, 420-425

29. Staden, R. (1985) in *Genetic Engineering: Principles and Methods* (Setlow, J. K., and Hollaender, A., eds) Vol. 7, pp. 67-114, Plenum Publishing Corp., New York

30. Staden, R. (1988) *Comput. Appl. Biosci.* 4, 53-60

31. Ratner, L., Fisher, A., Jagodziński, L. L., Mitsuya, H., Liou, R.-S., Gallo, R. C., and Wong-Staal, F. (1987) *AIDS Res. Hum. Retroviruses* 3, 57-69

32. Loeb, D. D., Hutchinson, C. A., III, Edgell, M. H., Farmerie, W. G., and Swanstrom, R. (1989) *J. Virol.* 63, 111-121

33. Loeb, D. D., Swanstrom, R., Everett, L., Manchester, M., Stamper, S. E., and Hutchison, C. A., III (1989) *Nature* 340, 397-400

34. Vieira, J., and Messing, J. (1987) *Methods Enzymol.* 153, 3-11

35. Kunkel, T. A., Roberts, J. D., and Zakour, R. A. (1987) *Methods Enzymol.* 153, 367-382

36. Sanger, F., Nicklen, S., and Coulson, A. R. (1977) *Proc. Natl. Acad. Sci. U. S. A.* 74, 5463-5467

37. Maniatis, T., Fritsch, E. F., and Sambrook, J. (1982) *Molecular Cloning: A Laboratory Manual*, pp. 1-545, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY

38. Laemmli, U. K. (1970) *Nature* 227, 680-686

39. Kyte, J., and Doolittle, R. F. (1982) *J. Mol. Biol.* 157, 105-132

40. Farmerie, W. G., Loeb, D. D., Cassevant, N. C., Hutchinson, C. A., III, Edgell, M. H., and Swanstrom, R. (1987) *Science* 236, 305-308

41. Pearl, L. H., and Taylor, W. R. (1987) *Nature* 328, 482

42. Oroszlan, S., and Copeland, T. D. (1985) *Curr. Top. Microbiol. Immunol.* 115, 221-233

43. Henderson, L. E., Benveniste, R. E., Bowder, R., Copeland, T. D., Schultz, A. M., and Oroszlan, S. (1988) *J. Virol.* 62, 2587-2595

44. Gowda, S. D., Stein, B. S., and Engleman, E. G. (1989) *J. Biol. Chem.* 264, 8459-8462

45. Mervis, R. J., Ahmad, N., Lillevoj, E. P., Raum, M. G., Salazar, F. H., Chan, H. W., and Venkatesan, S. (1988) *J. Virol.* 62, 3993-4002

46. Erickson-Viitanen, S., Manfredi, J., Viitanen, P., Tribe, D. E., Titch, R., Hutchinson, C. A., III, Loeb, D. D., and Swanstrom, R. (1989) *AIDS Res. Hum. Retroviruses* 5, 577-591

47. Krausslich, H. G., Ingraham, R. H., Skoog, M. T., Wimmer, E., Pallai, P. V., and Carter, C. A. (1989) *Proc. Natl. Acad. Sci. U. S. A.* 86, 807-811

48. Blackburn, S. (1978) in *Amino Acid Determination Methods and Techniques* (Blackburn, S., ed) pp. 7-37, Marcel Dekker Inc., New York

49. Davies, D. R. (1990) *Annu. Rev. Biophys. Biophys. Chem.* 19, 189-215

50. Richardson, J., and Richardson, D. (1988) in *Prediction of Protein Structure and Principles of Protein Conformation* (Fasman, G., ed) pp. 1-88, Plenum Publishing Corp., New York

51. Partin, K., Krausslich, H. G., Ehrlich, L., Wimmer, E., and Carter, C. (1990) *J. Virol.* 64, 3938-3947

52. Myers, G., Rabson, A. B., Berzofsky, J. A., Smith, T. F., and Wong-Staal, F. (1990) *Human Retroviruses and AIDS*, Los Alamos National Laboratory, Los Alamos, NM

53. Oroszlan, S., Copeland, T. D., Henderson, L. E., Stephenson, J. R., and Gilden, R. V. (1979) *Proc. Natl. Acad. Sci. U. S. A.* 76, 2996-3000

54. Copeland, T. D., Morgan, M. A., and Oroszlan, S. (1983) *FEBS Lett.* 156, 37-40

55. Yoshinaka, Y., Kato, I., Copeland, T. D., Smythers, G. W., and Oroszlan, S. (1986) *J. Virol.* 57, 826-832

56. Henderson, L. E., Bowder, R. C., Smythers, G. W., and Oroszlan, S. (1987) *J. Virol.* 61, 1116-1124

57. Devare, S. G., Kim, Y., Fox, J. D., Getchell, J., Cabradill, C., and Kalayanaraman, V. S. (1985) *Virology* 142, 206-210

58. Veromese, F. D., Copeland, T. D., Devico, A. L., Rehman, R., Oroszlan, S., Gallo, R. C., and Sarnagadarhan, M. G. (1986) *Science* 231, 1289-1291

59. Lightfoot, M. M., Coligan, J. E., Folks, T. M., Fauci, A. S., Martin, M. A., and Venkatesan, S. (1986) *J. Virol.* 60, 771-775

60. Hizi, A., Henderson, L. E., Copeland, T. D., Bowder, R. C., Krtzsch, H. C., and Oroszlan, S. (1989) *J. Virol.* 63, 2543-2549

61. Hizi, A., Henderson, L. E., Copeland, T. D., Bowder, D. C., Hixson, C. V., and Oroszlan, S. (1987) *Proc. Natl. Acad. Sci. U. S. A.* 84, 7041-7045

62. Henderson, L. E., Bowder, R., Smythers, G., Benveniste, R. E., and Oroszlan, S. (1985) *J. Virol.* 55, 778-787

63. Oroszlan, S., Henderson, L. E., Stephenson, J. R., Copeland, T. D., Long, C. W., Ihle, J. N., and Gilden, R. V. (1978) *Proc. Natl. Acad. Sci. U. S. A.* 75, 1404-1408

64. Copeland, T. D., Gerard, G. F., Hixson, C. W., and Oroszlan, S. (1985) *Virology* 143, 676-679

65. Henderson, L. E., Bowder, R., Copeland, T. D., Smythers, G., and Oroszlan, S. (1984) *J. Virol.* 52, 492-500

66. Vogt, V. M., Pepinsky, R. B., and Southard, L. E. (1985) *J. Virol.* 56, 31-39

67. Hunter, E., Bennett, J. C., Brawn, A., Pepinsky, R. B., and Voght, V. M. (1983) *J. Virol.* 45, 885-888

68. Brawn, A. S., Bennett, J. C., and Hunter, E. (1980) *J. Biol. Chem.* 255, 6962-6965

69. Misano, K. S., Sharief, F. S., and Leis, J. P. (1980) *Fed. Proc. Exp. Biol. Abs.* 60, 1611

70. Sauer, R. T., Allen, D. W., and Wali, H. D. (1981) *Biochemistry* 20, 3784-3791

71. Hippenmeyer, P. J., and Grandженетт, D. P. (1984) *Virology* 137, 358-370