Heterogeneity in recombinant HIV-1 integrase corrected by site-directed mutagenesis: the identification and elimination of a protease cleavage site

Alison B. Hickman, Fred Dyda and Robert Craigie

Laboratory of Molecular Biology, National Institute of Diabetes and Digestive and Kidney Diseases, National Institutes of Health, Bethesda, MD 20892, USA

1To whom correspondence should be addressed

Purified recombinant human immunodeficiency virus type 1 (HIV-1) integrase and certain deletion mutants exhibit heterogeneity consistent with proteolysis at a site close to the C-terminus. Electrospray ionization mass spectrometric analysis indicated that proteolytic cleavage generated a protein missing five residues from the C-terminus. PCR mutagenesis of amino acids on either side of the cleavage site identified two changes which were subsequently shown to prevent clipping when proteins were expressed and purified from *Escherichia coli*: the substitution of Arg284, the residue on the C-terminal side of the cleavage site, by either glycine or lysine. The introduction of either of these mutations into full-length integrase did not affect *in vitro* 3' processing or strand transfer activities. Thus, the incorporation of either of these mutations is likely to be beneficial when homogeneity of HIV-1 integrase is a concern, as in crystallographic or nuclear magnetic resonance spectroscopic experiments.

Keywords: *Escherichia coli*/HIV/integrase/protease/site-directed mutagenesis

**Introduction**

As part of the retroviral life cycle, virally encoded integrase removes two nucleotides from each 3' end of newly reverse transcribed viral DNA and then inserts the viral DNA into host cell target DNA (for reviews, see Vink and Plasterk, 1993; Katz and Skalka, 1994). This integration of viral DNA is required for the subsequent generation of new viral particles and is an obvious step to target in the effort to identify anti-HIV drugs. We are interested in the three-dimensional structure of HIV-1 integrase, both as a means of understanding the chemistry of catalysis and as a foundation for rational drug design. However, structural work on recombinantly expressed HIV-1 integrase has been impeded by its biophysical properties: low solubility and a proclivity for aggregation. A point mutation was recently described, F185K, which renders HIV-1 integrase (IN50–212; superscripts denote the amino acids comprising integrase or its domains) soluble and monodisperse at high protein concentrations (Jenkins *et al.*, 1995, 1996). This mutation allowed us to determine the structure of the catalytic domain of HIV-1 integrase, IN50–212, by X-ray crystallography (Dyda *et al.*, 1994). The structure of another fragment of integrase, IN220–270, a stretch of 50 amino acids located close to the C-terminus that possesses a non-specific DNA binding activity, has been determined using nuclear magnetic resonance (NMR) spectroscopic methods (Eijkelenboom *et al.*, 1995; Lodi *et al.*, 1995).

Despite the introduction of the F185K mutation into HIV-1 integrase, its high-resolution structure and information on the spatial organization of its functional domains continue to elude researchers. Although we have solved the problems of solubility and aggregation for HIV-1 integrase through mutation, there are myriad other reasons why a particular protein may not crystallize well. One of these is protein heterogeneity (McPherson, 1990). Heterogeneity can arise from many sources (oxidation, post-translational modification, etc.), or can result from contaminating proteins which are not removed during purification. We have observed that recombinantly expressed HIV-1 and HIV-2 integrase consistently co-purify with a slightly smaller protein (Figure 1, lanes 1 and 2). A smaller protein also co-purifies when integrase is expressed with a histidine tag (HT) at its N-terminus and purified by affinity chromatography. This suggests that the contaminating protein is a truncated form of integrase, rather than an *Escherichia coli* protein, an interpretation confirmed by N-terminal sequencing (Hickman *et al.*, 1994). The size of the contaminating protein also changes with the size of the integrase derivative (Figure 1, lane 4). The presence of a co-purifying protein correlates without exception with the extreme C-terminus of integrase. For example, preparations of the C-terminal domain alone, IN50–288, have substantial amounts of such a protein (Figure 1, lane 3), whereas C-terminal deleted constructs such as IN50–212 and IN1–212 do not exhibit this heterogeneity (Hickman *et al.*, 1994; unpublished data).

We hypothesized that, during expression in *E.coli*, integrase...
Removal of a cleavage site in HIV-1 integrase

Fig. 2. Electrospray ionization mass spectrum of IN\textsuperscript{50–288}/F185K/C280S from which the histidine tag had been removed by thrombin cleavage. This procedure leaves three non-integrase residues at the N-terminus, GSH–, which were included in calculations of molecular mass. (A) The original spectrum shows an envelope of peaks corresponding to different mass-to-charge (m/z) states. These peaks occur in pairs, with the larger of each pair (e.g. peaks with m/z 2245.5, 2073.1 and 1925.2) corresponding to the full-length protein and the smaller (peaks with m/z 2192.2, 2023.4 and 1879.1), presumably corresponding to the co-purifying truncated integrase. (B) Deconvoluted spectrum indicating two species. The major species is –HT IN\textsuperscript{50–288}/F185K/C280S bound to a single sodium ion, while the minor species corresponds to –HT IN\textsuperscript{50–283}/F185K/C280S, also bound to a single sodium ion.

Fig. 3. SDS–PAGE analysis of purified +HT IN\textsuperscript{50–288} with mutations at positions 283 or 284 as indicated (first six lanes). The wild-type protein corresponds to +HT HIV-1 IN\textsuperscript{50–288}/F185K/C280S.

is a substrate for a bacterial protease that either cuts the protein at a specific site close to the C-terminus, or exopeptidolytically removes a certain number of accessible residues. A truncated protein is seen by SDS–PAGE when cells expressing IN\textsuperscript{213–288} are lyzed directly in sample buffer, indicating that, at least in this case, proteolysis occurs during expression, not during subsequent purification. Proteolytic removal of amino acids is a concern as it has been reported that integrase missing five amino acids from its C-terminus is only partially active for strand transfer (Drellich et al., 1992), one of the activities of integrase that can be monitored in vitro using short oligonucleotide substrates that mimic the ends of viral DNA. Thus, these last few amino acids may be crucial contributors to enzyme activity, or may be required to maintain the three-dimensional structure of the protein. Although protein heterogeneity arising from proteolysis may not be the direct cause of our inability to obtain suitable integrase crystals, we undertook to determine the site of cleavage and devise methods to eliminate it.

Materials and methods

PCR-based mutagenesis

Amino acid changes were introduced at residue 283 or 284 and a stop codon introduced after residue 283, by PCR-based mutagenesis using plasmid DNA encoding IN\textsuperscript{1–283}/F185K/C280S as the template DNA and the appropriate primers (BioServe). The resulting PCR fragments were gel-purified, digested with NdeI and BamHI and then ligated with NdeI–BamHI-digested pET-15b (Novagen), which adds a 20 amino acid tag containing six histidine residues to the N-terminus of each expressed protein. The gene sequences of all mutated proteins were confirmed by DNA sequencing (New England Biolabs CircumVent Thermal Cycle Sequencing Kit).

Protein expression and purification

Expression of various integrase proteins in E.coli strain BL21(DE3) was as described previously (Craigie et al., 1995). Untagged, full-length HIV-1 integrase was purified by conventional chromatography as described previously (Craigie et al., 1995). Untagged, full-length HIV-2 integrase was extracted from the insoluble pellet and precipitated using ammonium sulfate as described by Craigie et al. (1995) for HIV-1 integrase. It was further purified by chromatography on phosphocellulose in 1.5 M urea-containing buffers (gradient elution from 0.1 to 1.0 M NaCl). Unless stated otherwise, all other integrase derivatives contain F185K and C280S mutations as appropriate. The first mutation, F185K, renders integrase and its deletion mutants soluble while the second prevents time-dependent aggregation in the absence of reducing agents (Jenkins et al., 1995, 1996).

Histidine-tagged (HT) proteins such as +HT IN\textsuperscript{50–288} and +HT IN\textsuperscript{1–288} were purified using Ni\textsuperscript{2+} affinity chromatography as described previously (Jenkins et al., 1996). To screen rapidly for proteolysis, mutated +HT IN\textsuperscript{50–288}, +HT IN\textsuperscript{213–288} and +HT IN\textsuperscript{217–288} derivatives were purified on a small scale by batch Ni\textsuperscript{2+} affinity chromatography as follows:

Induced cells (50 ml of culture induced at OD\textsubscript{600 nm} = 0.8) were resuspended in 1 ml of 25 mM HEPES, pH 7.5, 0.1 mM EDTA and frozen in liquid nitrogen. Cells were then thawed on ice and buffer was added to bring the solution to final concentrations of 0.3 mg/ml lysozyme, 0.5 M NaCl and 2 mM β-mercaptoethanol (β-ME). After 30 min on ice, lyzed cells were frozen in liquid nitrogen, allowed to thaw on ice and then centrifuged in a Beckman TL-100 ultracentrifuge at 100 000 g for 45 min at 4°C. (All subsequent steps were carried out at 4°C.) Chelating Sepharose (Pharmacia) was equilibrated with 50 mM NiSO\textsubscript{4} according to the manufacturer’s procedures and resuspended 1:1 (v/v) with buffer A (20 mM HEPES, pH 7.5, 0.5 M NaCl, 2 mM β-ME) containing 5 mM imidazole (Im). This suspension was added 1:1 (v/v) to the supernatant above and the mixture rotated slowly for 4 h to batch adsorb the protein on the resin. The resin was pelleted by low-speed centrifugation and the supernatant removed. The resin was washed with an equal volume of buffer A containing 5 mM Im by rotating again for 5 min and centrifuging. This was repeated 3–4 times. The resin was then washed 3–4 times as described above with buffer A containing 60 mM Im. The protein was subsequently eluted using a minimal volume of buffer A containing 0.8 M Im by rotating for 20 min. Imidazole was removed from the final preparation by dialysis against 20 mM Tris, pH 7.5, 1 M NaCl, 1 mM EDTA and 1 mM diithiothreitol (DTT). Purified proteins were analyzed for heterogeneity by SDS–PAGE and for enzymatic activity as described below. For the purification of +HT IN\textsuperscript{213–288} and +HT IN\textsuperscript{217–288} mutant proteins, β-ME and DTT were omitted.

Sample preparation for electrospray ionization mass spectrometry (ESI-MS)

The histidine tag (HT) was removed from purified +HT IN\textsuperscript{50–288} by thrombin cleavage as described previously (Craigie et al., 1995). The protein was dialyzed for >70 h against 5% acetic acid–0.1 mM DTT with several buffer changes and then for 24 h against 5% acetic acid alone. The extensive protein...
A.B. Hickman, F. Dyda and R. Craigie

Fig. 4. In vitro 3′ processing and strand transfer activities. (A) 3′-processing activities of indicated proteins over a range of protein concentrations (in µM) were assayed as described in the text. The substrate (S) and product (P) positions are indicated on the left. The band in all lanes which appears at the −1 position is a contaminant in the substrate. (B) Strand transfer activities of same proteins over same range of protein concentrations. The products (P) of strand transfer are a ladder of oligonucleotides of various sizes corresponding to different positions of integration. All assays were performed using +HT proteins.

Fig. 5. SDS–PAGE analysis of purified C-terminal fragments of HIV-1 integrase with amino acid changes introduced at Arg284. Lanes: 1, +HT IN213–288/C280S; 2, +HT IN217–288/C280S; 3, +HT IN217–288/C280S/R284G; 4, +HT IN217–288/C280S/R284K.

Fig. 6. Gel filtration elution profile of –HT IN217–288/C280S/R284G on a Superdex 200 column. High molecular mass material, typically aggregated protein, elutes in the column void volume at ~18 min.

Integrase enzyme assays

The 32P-labelled substrate was prepared by annealing 5′-ACTGCTAGAGATTTCCACAC (AE117) and 5′-GTGTGGAAAAATCTCTAGCAGT (AE118) as described previously (Engelman and Craigie, 1992). The standard reaction buffer (final volume 16 µl) contained 25 mM MOPS, pH 7.2, 10 mM β-ME, 0.1 mg/ml bovine serum albumin, 7.5 mM MnCl2, 10% glycerol, ~0.4 pmol substrate and various concentrations of protein as indicated in Figure 4A and B. Samples were incubated at 37°C for 60 min and the reactions stopped by the addition of 16 µl of sequencing gel loading dye. Samples were heated at 100°C for 3–5 min and 2 µl electrophoresed in 20% acrylamide–urea gels. The results were visualized by autoradiography.
Removal of a cleavage site in HIV-1 integrase

## Results and discussion

Electrospray ionization mass spectrometry (ESI-MS) was used to determine the molecular mass of the contaminant present in integrase preparations. A sample of –HT IN^{217-288}/R284G was dialyzed extensively into 5% acetic acid–0.1 mM DTT to remove interfering buffer components. The resulting ESI mass spectrum is shown in Figure 2. The series of paired mass-to-charge (m/z) peaks indicate that there is a predominant species and a minor one comprising ~5% of the total sample. The deconvoluted spectrum is shown in Figure 2B. The molecular mass of the major species (26 948 Da) corresponds to the intact protein (calculated molecular mass = 26 925 Da) bound to a sodium ion (23 Da). Assuming that the minor species (26 302 Da) also binds a single sodium ion, it corresponds to an integrase derivative missing five amino acids from its C-terminus, since –HT IN^{30-283} has a calculated molecular mass of 26 282 Da.

Inspection of the database of sequenced viral strains demonstrated that although the extreme C-terminus of HIV-1 integrase is highly conserved, the residues on either side of the putative site of proteolysis, Ser283 and Arg284, are substituted in viral strains, either singly or in combination, with glycine (Table I, A). These are substitutions, therefore, that do not eliminate in vivo integrase activity. Guided by this information, we created a series of IN^{1-288} and IN^{50-288} derivatives in which mutations were introduced at positions 283 or 284 (Table I, B).

To determine the extent of protein heterogeneity with different amino acids at positions 283 and 284, each +HT IN^{50-288} derivative was purified by batch Ni^{2+} affinity chromatography and analyzed by SDS–PAGE (Figure 3). Two of the mutations, R284K and R284G, yielded proteins which appeared as single bands, rather than the previously observed characteristic doublet. Identical results were obtained for the IN^{1-288} versions (data not shown). IN^{50-283} migrated at the same position as the contaminating protein in cases where cutting occurred, confirming our interpretation of the ESI-MS spectrum (compare the last two lanes in Figure 3).

To determine if the R284K or R284G changes affect in vitro enzyme activity, the 3’ processing and strand transfer activities of +HT IN^{1-288}, +HT IN^{1-289}/R284G, +HT IN^{1-289}/R284K and +HT IN^{213-283} were compared (Figure 4A and B). The first assay, 3’ processing, measures the ability of integrase to remove two nucleotides from the 3’ end of an oligonucleotide containing the U5 viral end sequence, while strand transfer measures the subsequent insertion of an oligonucleotide with a recessed 3’ end into a second oligonucleotide (Sherman and Fye, 1990; Craigie et al., 1991). In contrast to the results of Drelich et al. (1992), truncation at residue 283 did not lead to a measurable decrease in either Mn^{2+}-dependent 3’ processing or strand transfer activity. This is consistent with the observations of Vink et al. (1993) for HIV-1 integrase fused at its N-terminus to the maltose-binding protein, which showed no decrease in in vitro activity when nine amino acids were removed from its C-terminus. As might be expected from the natural variation in HIV-1 integrase sequences, +HT IN^{1-288}/R284G was as active in both assays as the unmutated version; this was also true for +HT IN^{1-288}/R284K. All activity assays were repeated using proteins from which the HT were removed by thrombin cleavage and the results were unaffected (data not shown).

The dramatic improvement of sample homogeneity resulting from either the R284G or R284K change is demonstrated in Figure 5, where a construct comprising the C-terminal domain alone, IN^{217-288}, was mutated to incorporate each change. Purified –HT IN^{217-288}/R284G was determined by analytical gel filtration to elute at a position consistent with a dimer (~70%), while IN^{1-288}/R284G, IN^{213-283}/R284G, and IN^{50-283}/R284G migrated at the same position as the contaminating protein in cases where cutting occurred, confirming our interpretation of the ESI-MS spectrum (compare the last two lanes in Figure 3).

In summary, we have introduced a mutation into HIV-1 integrase, substitution of residue 284 from either Gly or Lys, that prevents protein heterogeneity caused by a presumed E.coli protease. This mutation will be valuable in improving the homogeneity of our protein preparations for crystallographic studies, particularly in cases such as IN^{213-288}/R284G where at least 25% of the unmutated protein was proteolytically processed. We presume such an approach can be readily applied to other proteins that are clipped during expression or purification, since both PCR mutagenesis and the small-scale protein purification method described here are fairly rapid procedures. As two substitutions at position 284 were able to prevent the formation of a truncated protein, this indicates that the problematic E.coli protease is most likely an endoprotease, rather than an exoprotease which successively removes the last five amino acids and then stops. HIV-2 integrase, a protein that is also proteolytically processed when recombinantly

### Table I. C-terminal sequences of HIV-1 integrase

<table>
<thead>
<tr>
<th>Original sequence:</th>
<th>283</th>
<th>288</th>
</tr>
</thead>
<tbody>
<tr>
<td>–AGDDCV A S</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

(A) Naturally occurring mutants*

| –AGDDCV A G         |     |     |
| –AGDDCV A G QDE D   |     |     |

(B) Mutations made

| –AGDDVSA S H QDE D  |     |     |
| –AGDDVSA K QDE D    |     |     |
| –AGDDVSA A QDE D    |     |     |
| –AGDDVSA G R QDE D  |     |     |
| –AGDDVSA A R QDE D  |     |     |
| –AGDDVSA S A QDE D  |     |     |

* Naturally occurring mutants were identified using the NCBI (National Center for Biotechnology Information) BLAST program by searching available peptide sequence databases.

### Gel filtration of –HT IN^{217-288}/R284G

The C-terminal domain incorporating one of the mutations that prevents cleavage, +HT IN^{217-288}/R284G, was purified from the soluble fraction of lysed E.coli cells essentially as described for IN^{50-217}/F185K (Jenkins et al., 1995), except that reducing agents were omitted. After removal of the HT by thrombin cleavage, protein was concentrated to ~19 mg/ml using a Centriprep 10 (Amicon) and dialyzed against 1 M NaCl, 20 mM Tris, pH 7.5, 1 mM EDTA, 10% (w/v) glycerol.

Analytical gel filtration (sample size 15 µl) was performed on a Pharmacia SMART System using a pre-packed Superdex 200 PC 3.2/30 column; the column was equilibrated in dialysis buffer at 4°C prior to sample loading.
expressed (see Figure 1, lane 2), has an Ala–Arg sequence located close to its C-terminus (Arg is seven amino acids from end), reminiscent of the Ser283–Arg284 sequence in HIV-1 integrase. This suggests that similar mutations in this protein may also improve its homogeneity.

Our approach of site-directed mutagenesis is clearly not the only method to circumvent a protease problem during expression. For example, it is possible to express proteins in host E.coli strains defective for certain proteases (Das, 1990), although this may be a time-consuming and haphazard approach unless the sequence at the susceptible site is already known. This information may not even be sufficient as, although we have characterized the cleavage site in HIV integrase, we have been unable to identify an E.coli protease responsible for hydrolysis prior to an arginine residue. An alternative approach which was tried, moving the His tag to the C-terminus which should select during purification for molecules which are not cut at the C-terminus, was not successful for HIV-1 integrase as an entirely new set of proteolytic products was observed (T.Jenkins, personal communication).

Acknowledgements

We are grateful to Lewis K.Pannell, Laboratory of Analytical Chemistry, NIDDK, NIH, for obtaining the mass spectra shown in Figure 2. We thank A.Engelman and T.Jenkins for instruction and guidance in molecular biological techniques, K.Mizuuchi for helpful discussions and W.Gillette for comments on the manuscript. This work was supported in part by the National Institutes of Health Intramural AIDS Targeted Antiviral Program.

References


Received October 17, 1996; revised January 24, 1997; accepted January 28, 1997