THE JOURNAL OF BIOLOGICAL CHEMISTRY
Vol. 280, No. 12, Issue of March 25, pp. 11369–11378, 2005


# Folding Regulates Autoprocessing of HIV-1 Protease Precursor*[S]

Received for publication, November 8, 2004, and in revised form, December 23, 2004
Published, JBC Papers in Press, January 4, 2005, DOI 10.1074/jbc.M412603200

**Amarnath Chatterjee‡, P. Mridula‡, Ram Kumar Mishra§, Rohit Mittal§, and Ramakrishna V. Hosur‡¶**

*From the Departments of ‡Chemical Sciences and §Biological Sciences, Tata Institute of Fundamental Research, Homi Bhabha Road, Mumbai 400 005, India*

**Autoprocessing of HIV-1 protease (PR) precursors is a crucial step in the generation of the mature protease. Very little is known regarding the molecular mechanism and regulation of this important process in the viral life cycle. In this context we report here the first and complete residue level investigations on the structural and folding characteristics of the 17-kDa precursor TFR-PR-$C^{nn}$ (161 residues) of HIV-1 protease. The precursor shows autoprocessing activity indicating that the solution has a certain population of the folded active dimer. Removal of the 5-residue extension, $C^{nn}$ at the C-terminal of PR enhanced the activity to some extent. However, NMR structural characterization of the precursor containing a mutation, D25N in the PR at pH 5.2 and 32 °C under different conditions of partial and complete denaturation by urea, indicate that the precursor has a high tendency to be unfolded. The major population in the ensemble displays some weak folding propensities in both the TFR and the PR regions, and many of these in the PR region are the non-native type. As both D25N mutant and wild-type PR are known to fold efficiently to the same native dimeric form, we infer that TFR cleavage enables removal of the non-native type of preferences in the PR domain to cause constructive folding of the protein. These results indicate that intrinsic structural and folding preferences in the precursor would have important regulatory roles in the autoprocessing reaction and generation of the mature enzyme.**

Retroviruses including human immunodeficiency virus (HIV),[1] use their minimal genetic information by encoding their structural proteins and enzymes as two polyprotein precursors Gag and Gag-Pol (1). Autoprocessing of these precursors is an essential step in the life cycle of the virus (2, 3). In HIV, HIV-1 protease (PR) plays a crucial role in virus maturation by processing these precursors into functional proteins (4, 5). The HIV-1 PR is a 22-kDa homodimeric aspartyl protease, with each monomer having 99 amino acids and contributing the conserved catalytic sequence Asp-(Ser/Thr)-Gly (6–9). As the HIV-1 protease, which is flanked by the highly variable $p6^{pol}$ at its N terminus and by the reverse transcriptase (RT) at its C terminus (10, 11), is responsible for all cleavages in the Gag-Pol precursor, its dimerization and autocatalytic release from the Gag-Pol are critical steps in the viral life cycle (4, 13, 14). Earlier studies have shown that premature activation or partial inhibition of the protease leads to retarded viral maturation (15–18). Hence understanding the exact sequence of protease maturation from the Gag-Pol precursor has gained importance in recent years because of its intrinsic importance in viral maturation and as a target for drugs against AIDS (19).

Pettit *et al.* (20) have recently shown, by co-expressing equivalent amounts of substituted Gag-Pol constructs, that the initial cleavage of the HIV-1 Gag-Pol precursor is intramolecular. Moreover, they showed that competitive active site inhibition by the drug retonavir was 10,000-fold less for the protease embedded in the precursor than for the mature free protease (20). Earlier, kinetic studies on the model precursor system MBP-ΔTF-Protease-ΔRT showed that the protease maturation takes place in two steps. (ΔTF and ΔRT are short native sequences from the transframe protein and the reverse transcriptase, respectively. MBP stands for maltose-binding protein of *Escherichia coli* containing two native cleavage sites, $p6^{pol}$/PR at the N terminus and PR/RT at the C terminus.) The first step involves an intramolecular cleavage of the N terminus that is followed by intermolecular cleavage of the C terminus (19, 21). A relatively low $K_m$ for peptide substrates representing the p6*-PR (where p6* is TFP+$p6^{pol}$) cleavage site, compared with that for oligopeptides corresponding to other Gag or Pol cleavage sites (23) supports the view that the N-terminal cleavage is an early event in the proteolytic cascade. The activity of the protease-ΔRT was found to be nearly equal to that of the mature PR, though its conformational stability was much less than that of PR (19). However a 600-fold decrease in catalytic activity was seen in MBP-ΔTF-protease-ΔRT compared with mature PR (19, 21). Thus the flanking N terminus of the protease seems to have important consequences with maturation.

The N-terminal transframe region (TFR) consisting of a conserved N-terminal transframe octapeptide (TFP) and a 48–60 amino acid long variable $p6^{pol}$, with a protease cleavage site at the intersection, does not have any stable secondary or tertiary structure in free solution (24), though some tendency for helix formation has been seen. However, when present with the PR, TFR does seem to act as a regulator for the autoprocessing of the protease (11, 23, 25–28, 29). Interaction of recombinant p6* protein with HIV-1 PR was found to specifically inhibit its activity, and the inhibition was dependent on the C-terminal cleavage site residues SFNF in the p6*. In separate experiments with the precursor, these residues blocked the substrate

* The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

[S] The on-line version of this article (available at http://www.jbc.org) contains Supplementary Materials.

¶ To whom correspondence should be addressed: Dept. of Chemical Sciences, Tata Institute of Fundamental Research, Homi Bhabha Road, Mumbai 400 005, India. Tel.: 91-22-2280-4545, extension: 2488; Fax: 91-22-2280-4610; E-mail: hosur@tifr.res.in.

[1] The abbreviations used are: HIV, human immunodeficiency virus; RT, reverse transcriptase; PR, HIV-1 protease; MBP, maltose-binding protein; TFR, N-terminal transframe region; AIDS, acquired immunodeficiency syndrome; MALDI-TOF, matrix-associated laser desorption ionization-time of flight; HSQC, heteronuclear single quantum coherence; TOCSY, total correlated spectroscopy; RT, reverse transcriptase; TFP, transframe octapeptide.

binding cleft in HIV1-PR after N-terminal autoprocessing of the precursor. At the same time it was also observed that the p6* stabilized the dimer, as the relative amount of dimer increased by 12% in its presence (25). Functional characterization of the model precursor ΔTFP-p6$^{pol}$-PR (ΔTFP is a 5-residue variant of TFP) by examination of the mechanism and the pH rate profile of the autocatalytic reaction to produce mature PR shows that full-length TFR with its native cleavage sites is critical for the regulated autoprocessing of Gag-Pol and for optimal catalytic activity (28). The extensive study by Dautin *et al.* (27) on functional modulations due to N- and C-terminal extensions to PR, using an *E. coli* genetic assay for proteolytic activity and a bacterial two-hybrid system, shows that the TFR can restore enzymatic activity to a dimerization-deficient HIV protease variant. Experiments with various deletion and addition mutants of PR and its precursors, Gag-Pol, TFR-PR, also give insights into folding and dimerization of PR (29). For example, deletion of the first four residues in PR led to >90% unfolded ΔPR. Similar destabilization was observed for PR with additional residues in the N terminus (29). Earlier, it has also been shown that removal of the p6$^{pol}$ domain from the Gag-Pol polyprotein leads to a significantly higher rate of processing of the Gag-ΔPol precursor (31).

The studies discussed so far are mainly based on enzymatic activity assays for the HIV-1 PR and its precursors using the chromogenic peptide substrate Lys-Ala-Arg-Val-Nle-Phe(p-$NO_2$)-Glu-Ala-Nle-$NH_2$ (19, 21, 32), or immunoblotting assays of the autolytic products. These give very good quantitative as well as qualitative information with regard to the working of the various precursors of HIV-1 protease. However, there are very few reports about the residue level structural characteristics of these precursors, which is crucial to understanding the molecular mechanism of protease maturation (29, 33). Louis *et al.* (29) have earlier shown, through NMR, how the N-terminal TFR extension to the HIV-1 PR does not allow it to fold even in the presence of DMP323, which is one of the tightest binding inhibitors. Detailed NMR structural characterization of wild-type TFP-p6$^{pol}$-PR was not possible because of its autolytic property. Hence, in a later study, an active site D25N mutation was introduced, and the HSQC spectra were seen to have many peaks at the same chemical shifts as in the spectra of the folded $PR_{D25N}$, though, they also had many intense peaks in a narrow region of amide proton chemical shifts (8.0–8.5 ppm), presumably belonging to the TFR residues. This indicated that the PR region folded properly, although the TFR region could not be characterized because of insufficient dispersion of the peaks (33). It was suggested that the TFR region was largely unstructured.

Thus, all the above studies demonstrate the importance of TFR on the folding and maturation of the protease. However, the mechanistic details at the residue level are still not understood. In this context we present here investigations on a precursor TFP-p6$^{pol}$-PR- C$^{nn}$, where C$^{nn}$ is a non-native pentapeptide extension at the C terminus of PR. Bacterial expression and MALDI analysis of the precursor show that TFR does not hamper the autoprocessing of the precursor so as to release the PR. Deletion of C$^{nn}$ enhanced autoprocessing, indicating that the non-native C-terminal extension interferes in the cleavage mechanism. We carried out extensive NMR investigations on the precursor containing an active site mutation D25N, which was stable for several weeks for NMR experiments. We monitored the intrinsic folding propensities of the precursor by studying the graded changes in the dynamic as well as structural characteristics of the equilibrium intermediates, created by use of different concentrations of the chemical denaturant, urea. These results have significant implications for the regulation mechanism of the autoprocessing reaction of HIV-1 protease precursors.

## MATERIALS AND METHODS

*Protein Preparation*—Starting with the clone for the TFR-PR-tethered dimer (TFR-PR-C$^{nn}$-PR), kindly supplied by Dr. M. V. Hosur of Bhabha Atomic Research Centre, Mumbai, we introduced an active site mutation, D25N, in PR, using a standard PCR-based site-directed mutagenesis strategy; this mutation does not affect PR folding but prevents its autocleavage. From this the TFR-PR-C$^{nn}$ region was selected and introduced into the NdeI/BamH1 multiple cloning site of a pET11a plasmid. The inclusion of the C$^{nn}$, besides providing a non-native flanking C terminus, has a practical advantage. It has the sequence GGSSG, and the glycines have a special significance in the NMR assignment strategy. At the same time, C$^{nn}$ is known not to affect the folding characteristics of PR in the tethered dimer, which folds similarly to the native homodimer (34). Similarly, the TFR-PR construct was also prepared. The desired wild-type constructs were prepared by PCR amplification of TFR-PR and TFR-PR-C$^{nn}$ regions from the full clone (TFR-PR-C$^{nn}$-PR) and introducing them into a pET11a plasmid as described above. The constructs were sequenced to verify that there were no inadvertent PCR-induced errors. The plasmid was transformed into *E. coli* strain BL21(DE3) for protein overexpression. Transformed bacteria were grown at 37 °C in M9 medium to $OD_{600}$ of ~0.8, and then induced for production of the desired proteins using 1 mM isopropyl-1-thio-β-D-galactopyranoside. Uniformly $^{15}N$- and $^{15}N/^{13}C$-labeled protein samples were prepared by growing bacteria in M9 minimal media supplemented with 1 g liter$^{-1}$ $^{15}NH_4Cl$ and 4 g liter$^{-1}$ [$^{13}C$]glucose. Protein was purified as described previously (35). MALDI analysis of the protein showed peaks at the expected molecular mass (17.3 kDa). The NMR samples contained 1 mM protein in 50 mM acetate buffer (pH 5.2) containing 5 mM EDTA, 20 mM dithiothreitol, and different concentrations of urea in 90% $H_2O$, 10% $D_2O$.

*Gel Electrophoresis*—The recombinant protein was induced in BL21(DE3) *E. coli* bacterial cells as described in the section on protein preparation. Aliquots were taken at two different induction times, 3 and 5 h, and analyzed on 12% SDS-PAGE.

*Capillary Electrophoresis*—The purified protein was concentrated to ~1 mM and analyzed by neutral capillary electrophoresis on a Beckmann-Coulter capillary electrophoresis system in the presence and absence of the denaturants, urea and guanidine hydrochloride.

*Mass Spectroscopy*—MALDI-TOF mass spectrometry analyses were carried out with Micromass (UK) MALDI-TOF Spec 2E spectrometer equipped with a UV nitrogen laser (337 nm) and a dual microchannel microplate detector. The samples were prepared by mixing 1 μl of protein solution (~20 μM) with 1 μl of freshly prepared matrix solution (10 mg/ml of 2,5-dihydroxybenzoic acid in 3:2 0.1% trifluoroacetic acid/acetonitrile). A total of 1 μl of this mixture was placed on the stainless steel probe plate and allowed to dry at room temperature. The spectra were recorded in the positive reflector linear mode at an accelerated voltage of 20 kV in the range from 4000 to 30,000 Da. For each measurement, the spectra were externally calibrated using myoglobin and trypsinogen.

*NMR Spectroscopy*—All NMR experiments were performed at 32 °C on a Varian Unity-plus 600 MHz NMR spectrometer equipped with pulse-shaping and pulse-field gradient capabilities. For the HNN spectrum the delays $T_N$, and $T_C$, were both set to 28 ms. 40 complex points were used along the $t_1$ and $t_2$ dimensions. The HN(C)N spectrum was recorded with the same $T_N$ and $T_C$ parameters, the same number of $t_1$ and $t_2$ points, and the $T_{CC}$ delay was set to 9 ms. TOCSY-HSQC was recorded with a mixing time of 60 ms, 32 complex points along the $^{15}N$ ($t_1$) dimension and 64 complex points along the $^1H$ ($t_2$) dimension. CBCANH and CBCA(CO)NH were recorded with 40 complex $t_1$ points ($^{15}N$) and 64 complex $t_2$ points ($^{13}C$). HNCO was recorded with 40 complex points along $t_1$ and $t_2$. An HSQC was recorded with 256 $t_1$ increments. For the high resolution HSQC data, required for coupling constant measurements, 8192 and 512 complex points were acquired along the $t_2$ and $t_1$ dimensions, respectively. For the relaxation measurements 2048 and 256 complex points were collected along the two dimensions. For $R_2$ measurements, the following Carr-Purcell-Meiboom-Gill (CPMG) delays were used: 10, 30, 50, 90, 130, 150, 190, 230 ms and spectra duplicated at 50 and 150 ms. The $R_2$ values were extracted by fitting the peak intensities to the equation $I(t) = B \exp(-R_2 t)$. The experiments were carried out using the pulse sequences described by Farrow *et al.* (36).

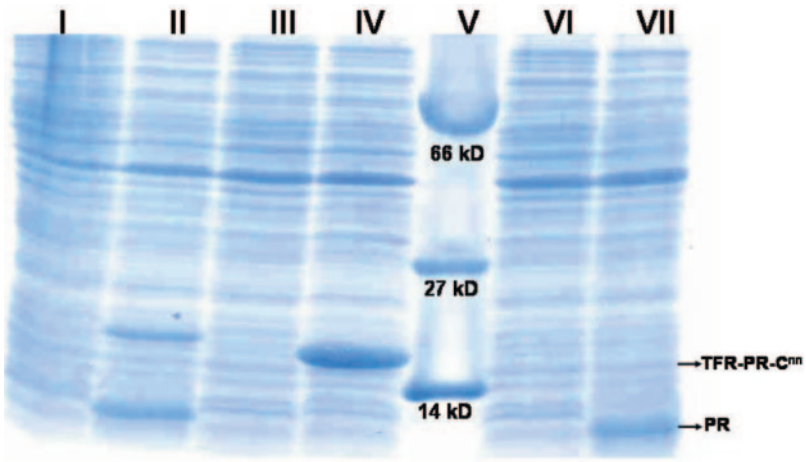


FIG. 1. **SDS-PAGE analysis of precursor activity.** *Lane I*, uninduced TFR-PR-$C^{nn}$; *lane II*, induced TFR-PR-$C^{nn}$; *lane III*, uninduced TFR-$PR^{D25N}$-$C^{nn}$; *lane IV*, induced TFR-$PR^{D25N}$-$C^{nn}$; *lane V*, molecular mass marker; *lane VI*, uninduced TFR-PR; *lane VII*, induced TFR-PR.

## RESULTS AND DISCUSSION

*C-terminal Extension at PR Retards the Autoprocessing Activity of the Precursor*—We checked the autoproteolytic activity of TFR-PR and TFR-PR-$C^{nn}$ precursors *in vivo* (Fig. 1, *lanes II* and *VII*). For TFR-PR, we see no trace of the precursor in the SDS-PAGE after 6 h of induction. This is clear evidence that the TFR does not prevent the autoprocessing activity of the precursor (*lane VII*). However, in the case of TFR-PR-$C^{nn}$ we see the presence of the intact precursor in the gel (*lane II*). Thus the C-terminal extension in our TFR-PR precursor slowed down autoprocessing. Our MALDI result with the purified protein also points to the same fact (Fig. 2). For the TFR-PR we see only an ~11-kDa peak for the PR and a ~7-kDa peak for the TFR part; however for the TFR-PR-$C^{nn}$ we see a peak at ~18 kDa corresponding to the precursor. This seems to suggest that the C-terminal extension possibly interacts with the PR region; either it interferes with dimer formation or it blocks the active site as has been observed for the SFNF stretch at the C terminus of the TFR in an earlier study (26).

*Intrinsic Folding Characteristics of the Precursor*—Since for the autocleavage reaction, the precursor has to become active by forming a dimer with the correct fold and generate an active site, we attempted to determine the structure of the precursor by NMR in solution. For this purpose we first prepared a D25N mutant of the precursor, which is inactive as the mutation is at the active site of PR. At the same time it is also known that the D25N mutation does not affect the folding of the protease (37). This mutant precursor is thus stable for weeks together and is ideally suited for structural characterization by NMR. However, it turned out that the protein had a high tendency to aggregate, as seen by dynamic light scattering, capillary electrophoresis (data not shown), and also by NMR (see below), over a wide range of experimental conditions of pH and temperature. Deleting the C-terminal extension also did not make any difference with regard to this behavior. This is at variance with the earlier report by Ishima *et al.* (33) on a precursor, which had the mutations Q7K, D25N, L33I, L63I, C67A in the PR region, and three residues at the N terminus of TFR that were different from those in our precursor. Ishima *et al.* (33) found the protein to be a monomer and stable even at the high NMR concentration, and the spectra displayed features of native-like fold for the PR region. Therefore to investigate the intrinsic folding characteristics of our present precursor, we undertook to elucidate the structural characteristics in 8 M urea and the transitions therefrom by NMR, and the various equilibrium intermediates were created by systematically varying the urea concentration. In the following, we first describe the sequence-specific resonance assignments for the various urea denatured states and then present the structural and dynamic characterizations of the precursor at pH 5.2 and 32 °C.

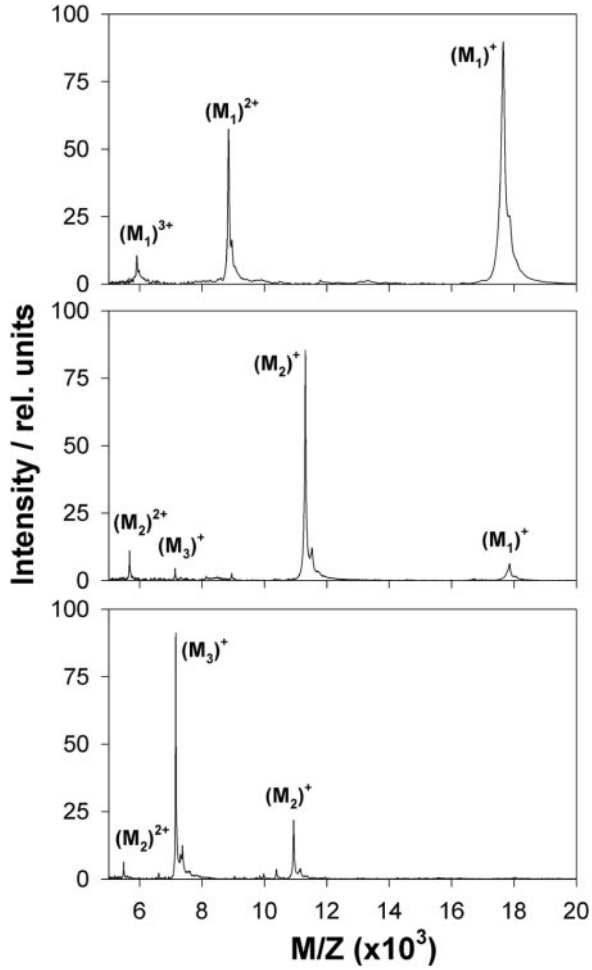


FIG. 2. **MALDI-TOF analysis of precursor activity.** *Top panel*, MALDI spectrum of TFR-PR-$C^{nn}$ containing the D25N active site mutation. Peaks attributed to singly $[M_1]^+$, doubly $[M_1]^{2+}$, and triply $[M_1]^{3+}$ charged species of the precursor are seen. *Middle panel*, MALDI spectrum of the precursor without the D25N mutation. The fragment peaks $[M_2]^+$ and $[M_3]^+$corresponding to the monomer PR and TFR are seen. A small peak corresponding to $[M_1]^+$ is also visible. *Bottom panel*, MALDI spectrum of the precursor TFR-PR. The intact precursor peak $[M_1]^+$ is not seen. Only the fragment peaks $[M_2]^+$ and $[M_3]^+$ are seen. In both the *middle* and the *bottom* panels doubly charged species of PR are also seen.

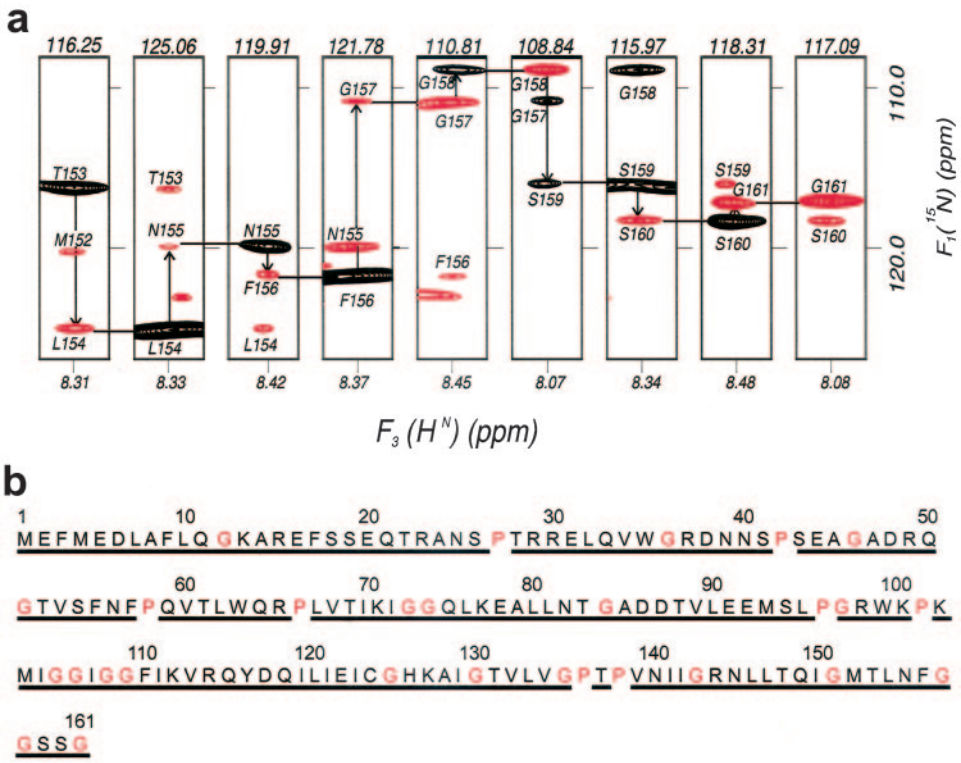


FIG. 3. **Sequential assignment of $H^N$ and $^{15}N$ from the HNN spectrum of the inactive precursor in 8 M urea.** *a*, $F_1$-$F_3$ strips at specific $F_2$ positions are aligned, and the sequential walk for the stretch 155–161 is shown as an illustration. *Black* and *red* contours indicate positive and negative peaks respectively. *b*, summary of all the observed sequential connectivities for the protein in the HNN spectrum.

*Resonance Assignments*—The TFR-PR-$C^{nn}$ precursor is 161 residues long, of which the first 57 residues belong to the TFR portion. The next 99 residues, that is, 58–156 actually constitute the PR portion. Hence residues 59–62, 152–156 form the dimerization domain, 82–84 form the active site, 138–140 form the substrate binding cleft, 100–106 form the hinge region, and 109–112 constitute the mobile flaps in the PR (35). The final five residues (157–161) having the sequence GGSSG, constitute an extension to the PR at the C terminus. Henceforth we will use these numbers for structural discussion.

Conventionally, backbone assignment in proteins has been achieved by a combination of several three-dimensional triple resonance experiments, typically, HNCA, HN(CO)CA, CBCANH, and CBCA(CO)NH (reviewed recently in Ref. 38). These experiments display correlations between $H^N$, $^{15}N$, and ($C^\alpha$, $C^b$) nuclei along the protein backbone. The success of this approach depends critically on the dispersion of the $C^\alpha$, $C^\beta$ chemical shifts, and therefore for unfolded proteins, where this dispersion is very poor, the method has serious limitations. Our methodology of assignment is based on the recently described triple resonance experiments HNN and HN(C)N (39). The most significant feature of these experiments is the observation of different patterns of positive and negative peaks in the ($F_1$, $F_3$) planes depending on the residue types at $i-1$, $i$, and $i+1$ positions. These have been discussed in detail earlier (39, 40); suffice it to say here that glycines and prolines play important roles in this regard, the former because of the absence of the $C^\beta$, and the latter because of the absence of the amide proton. Triplets containing these residues produce very characteristic patterns in the ($F_1$, $F_3$) planes, which can be termed as fixed points. These provide many starts and check points for the sequential walk, and hence it is less crucial to obtain side chain assignment to validate the backbone assignments. Nevertheless, simultaneous analysis of an $^{15}N$ resolved TOCSY (41) experiment helps in resolving occasional ambiguities in the connections because of degeneracies of the chemical shifts. This is particularly useful in unfolded proteins, since the side chain chemical shifts are close to their random coil values, and hence the spin systems of the residues can be relatively easily identified.

TFR-PR-$C^{nn}$ has 20 glycines and 8 prolines, which are well distributed over the length of the polypeptide chain. Thus there are a number of fixed points, well distributed, to enable unambiguous assignments. Fig. 3*A* shows an illustrative sequential walk through the stretch 153–161, and Fig. 3*B* displays the summary of the connectivities. All the amide and $^{15}N$ assignments are shown in the $^{15}N$ HSQC spectrum in Fig. 4.

Following the amide and $^{15}N$ assignments as discussed above the carbon assignments were readily obtained from the well known triple resonance experiments, CBCANH, CBCA(CO)NH, and HNCO (42, 43). The former two experiments together provide $C^\alpha$, $C^\beta$ assignments while the HNCO provides $C'$ assignments. We also obtained many side chain assignments for the individual residues from TOCSY-HSQC spectra in a straightforward manner, making use of the amide and $^{15}N$ assignments.

All the assignments made in 8 M urea have been listed in Table I of the Supplementary Material. The HSQC spectra at other urea concentrations were very similar to the one at 8 M urea (see below), and thus peak assignments could be readily obtained by simple transfer of assignments.

*Residual Structure at 8 M Urea*—It is now becoming increasingly evident that the denatured states of proteins are not always entirely random coils, but may contain regions with preferred conformations or propensities for transient structure formations (44–47). The regions having propensities for definite structures are the so-called folding cores, which indicate folding initiation sites on initiation of the folding reaction by dilution of the denaturant concentrations. We have probed for the existence of such preferences in the 8 M urea denatured

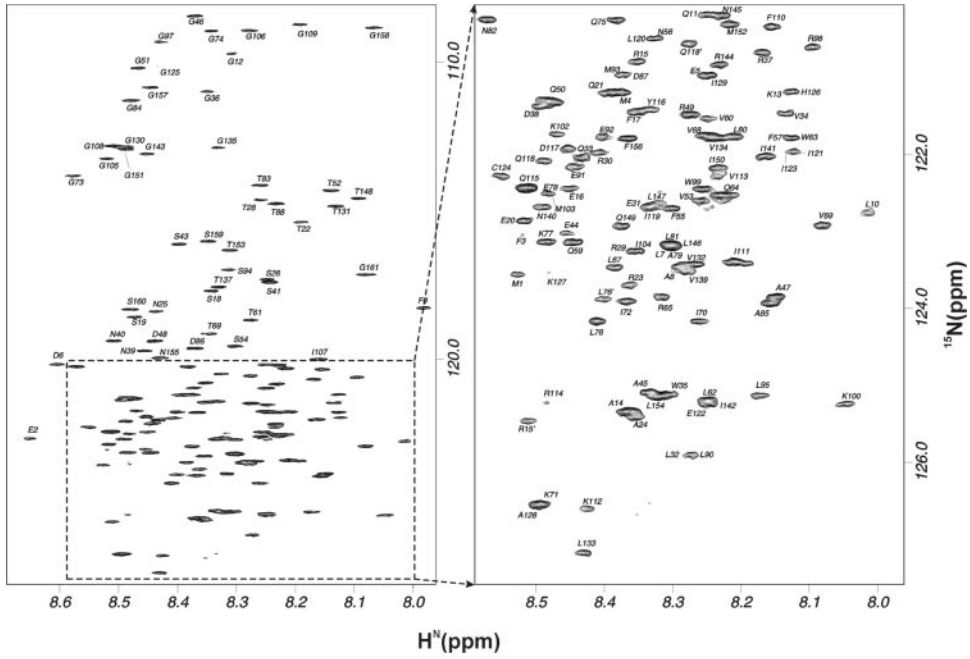


FIG. 4. **Finger print $^{1}$H-$^{15}$N HSQC spectrum in 8 M urea of the precursor TFR-PR-C$^{nn}$ containing the D25N mutation in the PR region, showing the residue-specific assignments at pH 5.2 and 32 °C.**

state of TFR-PR by using carbon C$^{\alpha}$, C′ secondary shifts (deviations of chemical shifts from their random coil values); these are believed to be the most diagnostic from the point of view of residual structural characterization (47, 48). Positive secondary shifts for $^{13}$C$^{\alpha}$ and $^{13}$C′ indicate a preference for $\varphi$, $\psi$ angles in the helical conformation, while negative secondary shifts indicate a preference for $\varphi$, $\psi$ angles in the $\beta$-sheet conformation. If a contiguous stretch of 3–4 residues shows a specific pattern of secondary shifts, that can be taken to indicate the presence of a transient secondary structural propensity in that region of the protein. Now, in any protein the observed chemical shifts are influenced both by neighboring amino acids and local backbone structure. Therefore, it is important to correct these for contributions from the local amino acid sequence (49). In the present analysis, the random coil values were corrected using sequence-dependent correction factors determined for a set of Ac-GGXGG-NH$_2$ peptides in 8 M urea and pH 2.3 (50). For the residues D, E, and H, which are sensitive to pH, the random coil values given by Wishart *et al.* (51) at pH 5.0, appropriately corrected for the alanine neighbor were used. Deviations in specific chemical shifts were then calculated by subtracting the corrected random coil values from the measured chemical shifts for all the residues in urea-unfolded TFR-PR-C$^{nn}$. These secondary shifts are shown in Fig. 5. The data does not seem to indicate the presence of any long stretches of preferred conformations but suggests the presence of many short contiguous regions with trends of $\alpha$, $\beta$ preferences.

The C$^{\alpha}$ secondary shifts (*top panel*) are rather small for most residues, but show interesting sequence-dependent variations; about 15 discrete residues show large secondary shifts (>1.5 ppm) which may represent individual preferences in the $\varphi$, $\psi$ space. There are many short contiguous stretches showing shifts of more than about 0.3 ppm (identified by the horizontal line in the figure), and these can be considered to be having propensities for transient secondary structure formation. There are also few other contiguous stretches with smaller shifts, which may indicate smaller tendencies to populate the respective $\varphi$, $\psi$ angles in the Ramachandran map. The locations of the stretches with good $\alpha$-helical and $\beta$-strand propensities are shown by open and filled boxes, respectively, in the figure.

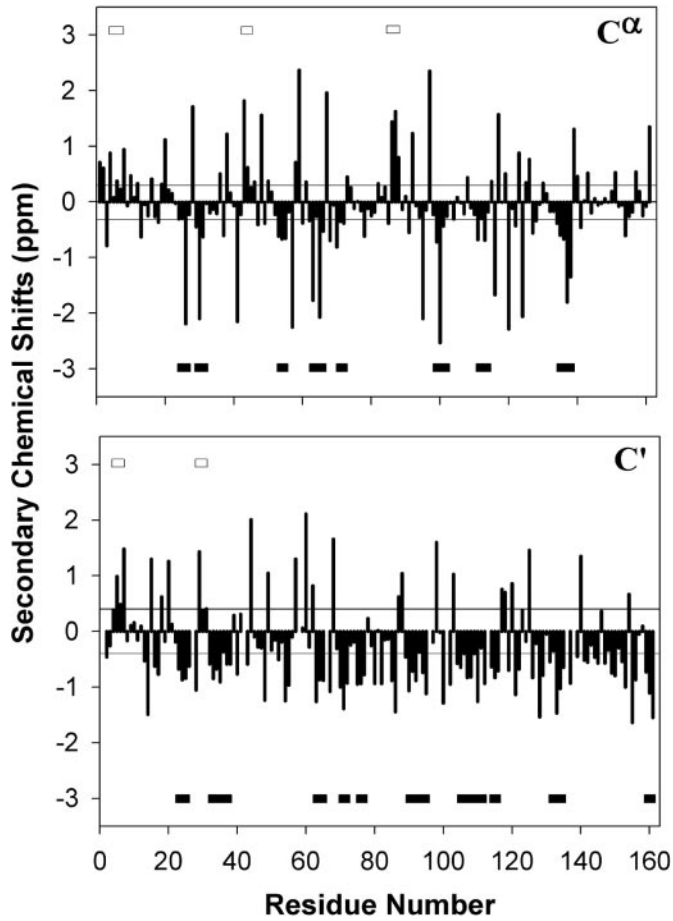


FIG. 5. **Sequence-corrected secondary chemical shifts.** Deviations of observed chemical shifts from sequence-corrected random coil values for C$^{\alpha}$ (*top panel*) and C′ (*bottom panel*) have been plotted against the residue number for the precursor in 8 M urea, pH 5.2 and 32 °C. Intrinsic propensities for $\beta$-strand and $\alpha$-helix formation are indicated by *filled* and *open boxes*, respectively, if at least three contiguous residues exhibit shifts larger than 0.3 ppm for C$^{\alpha}$ and 0.4 ppm for C′.

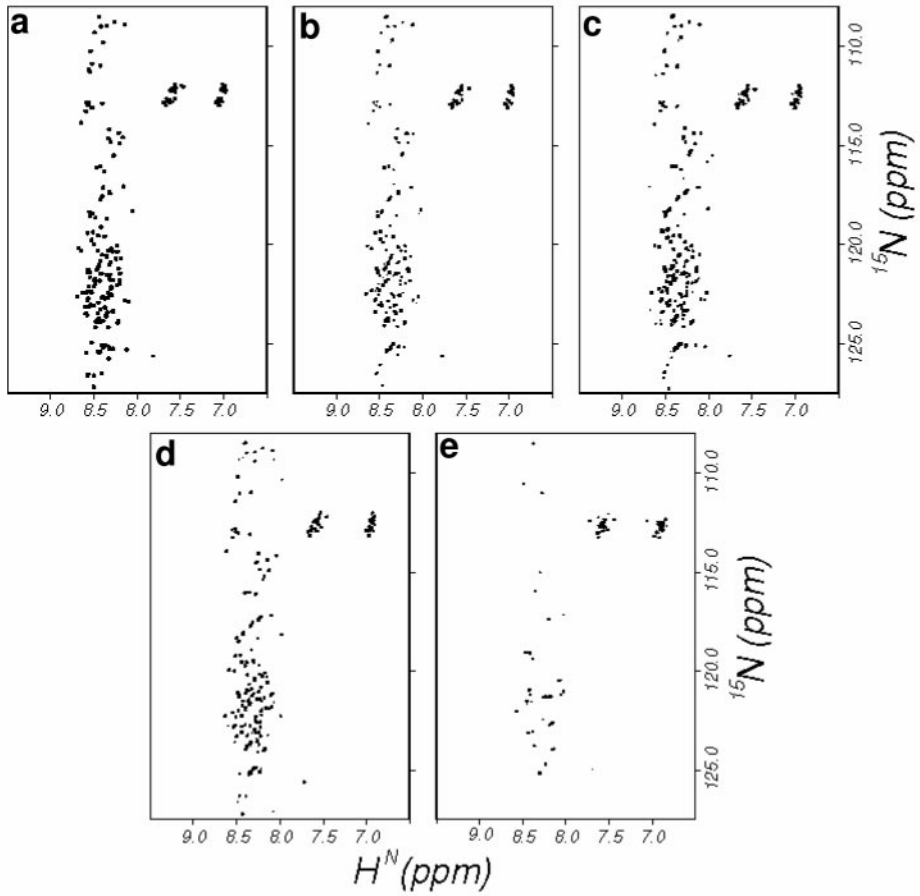


FIG. 6. **$^1$H-$^{15}$N HSQC spectra of the precursor with D25N mutation in the PR region at various urea concentrations, pH 5.2 and 32 °C.** *Panel a*, 8 M; *panel b*, 6 M; *panel c*, 4 M; *panel d*, 2 M; *panel e*, 0 M urea.

The TFR segment (residues 1–57) of the protein appears to contain two short $\beta$ segments and two short $\alpha$ segments. Previous qualitative reports on structural characteristics of TFR-PR (24, 28) suggested that the TFR segment may be largely unfolded in aqueous solutions. Our present observations, however, seem to suggest that there may be at least a few regions of some $\varphi$, $\psi$ preferences, in an otherwise largely unstructured polypeptide. The PR segment of the protein contains many $\beta$ segments and only one $\alpha$ stretch. The location of the $\alpha$ stretch (85–87) is certainly not the same as in the native PR where it occurs near the C terminus; this corresponds to the stretch 143–151 in the present case. Several of the $\beta$ stretches, namely, 62–66, 70–72, 98–102, 110–114, 134–138 belong to the native-type structures ($\beta$ type) in the dimeric structure of PR (34).

The C′ secondary shifts (*bottom panel* in Fig. 5) corroborate the results from $C^\alpha$ secondary shifts to a large extent. In both cases the contiguous stretches with $\alpha$ and $\beta$ propensities are nearly at the same locations as is also the discrete residues with large secondary shifts. The short displacements of the stretches or a few mismatches may be attributed to the facts that the segments themselves are very short, and the sensitivities of the C′ and $C^\alpha$ secondary shifts to $\alpha$, $\beta$ preferences could be slightly different. Overall, the C′ secondary shifts are slightly larger in magnitude compared with the $C^\alpha$ secondary shifts (cutoff of 0.4 ppm is used for C′ secondary shifts). The stretch at 103–111 is significantly longer, and this belongs to the flap segment of the native protease structure (34).

We also measured the $H^N$-$H^\alpha$ coupling constants (see below), amide proton temperature coefficients, and $^1$H-$^1$H nuclear overhauser effects (data not shown), all of which indicate that the protein is devoid of any persistent structure in 8 M urea. The sensitivities of the average coupling constants to the structural preferences are perhaps relatively smaller compared with the secondary chemical shifts. The transverse relaxation rates ($R_2$) (see below) indicated, however, sequence-dependent variations, suggesting possibilities of conformational transitions at certain locations. Thus we conclude that in 8 M urea at pH 5.2 and 32 °C, the polypeptide is largely unstructured but with short pockets of specific secondary structural propensities in a dynamic ensemble.

*Equilibrium Intermediates Along the Folding Funnel*—Equilibrium intermediates created by different denaturant concentrations help to understand the folding transitions along the folding pathway of a protein. Fig. 6 shows the HSQC spectra of the precursor as a function of denaturant concentration. Interestingly, the spectra (Fig. 6, *panels a–d*) at 8, 6, 4, and 2 M do not show any substantial change in the profile of peak dispersions, thus showing that the protein has a high tendency to be unfolded. All the peaks present in the 8 M spectrum are also present in the 6, 4, and 2 M spectra at almost identical positions, barring a few that show small shifts. However, there are some weaker peaks in the spectra at all the denaturant concentrations, which suggest the presence of other conformations that may be partially folded forms. The presence of these peaks indicates that the state identified by the conserved peaks in the spectra would have differences in the dynamic characteristics under the different denaturant conditions. The spectrum (Fig. 6, *panel e*) in the absence of urea shows very few broad peaks, which is consistent with the aggregation behavior of the protein discussed earlier.

*Folding Propensities*—We have characterized the structural transitions from the 8 M urea state to the various other urea-

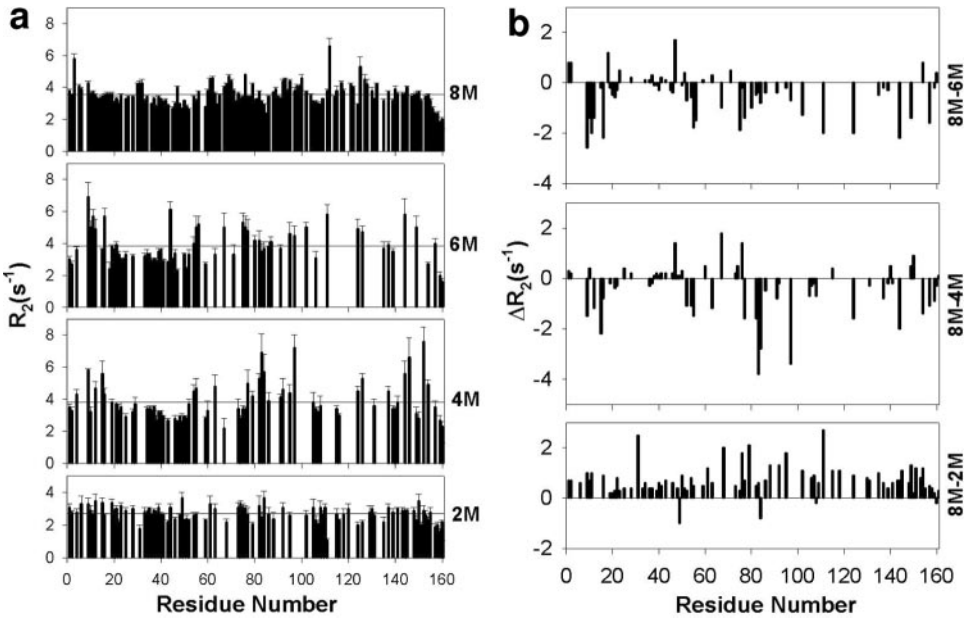


FIG. 7. **Transverse relaxation rates in the precursor.** *Left panel*, measured $R_2$ values at 8, 6, 4, and 2 M urea concentrations are plotted against the protein sequence. *Horizontal lines* in each box indicate average values. *Right panel*, deviations in the $R_2$ values at 6, 4, and 2 M urea from those at 8 M urea. *Horizontal bars* at the *bottom* of each panel show contiguous stretches with deviation of more than 1.5 $s^{-1}$.

created intermediates in our precursor using transverse relaxation rates ($R_2$) and three bond $H^N$-$H^\alpha$ coupling constants.

The transverse relaxation rates are sensitive to slow motions and conformational transitions occurring on the milli- to- microsecond time scale. In many instances these have provided valuable insights into sequence-dependent motional restrictions and flexibilities in denatured proteins, which in turn provide clues to the folding mechanisms (12, 52–57). In an earlier report, Bhavesh *et al.* (52) have shown the importance of sequence-specific variations in the transverse relaxation rates ($R_2$), as denaturant concentration is decreased, on the folding hierarchy of HIV-1 protease. The changes in the magnitude of $R_2$ values as the denaturant concentration is varied directly reflect on the transient conformational changes along the sequence that may lead to order, and hence native structure development by formation of native contacts or breaking of non-native contacts. Fig. 7*a* shows the $R_2$ values for the TFR-PR-$C^{nn}$ as the denaturant concentration is decreased from 8 to 2 M. The $R_2$ values do show sequence-specific variations indicating different degrees of restricted motions along the chain. Both upward and downward changes occur, suggesting sequence-dependent transient changes in the structural preferences. At this juncture we may mention that the absence of data points for some of the residues is caused by the difficulty in quantitation because of nearby weaker peaks, and also, data points having more than 15% fitting error have not been included; in most cases the errors are less than 6%. Fig. 7*b* shows the changes in the $R_2$ values as we move toward lower denaturant concentration. Negative and positive deviations indicate increase or decrease in $R_2$, respectively, and correspondingly represent increased and decreased conformational transitions, as long as the protein is still largely unfolded and there are no rigid structures. Once the rigid structures are formed, changes in $R_2$ values would be dictated by internal motions only. The deviations in Fig. 7*b* may be divided into three classes $\pm(>1.5)$, $\pm(1.5–1.0)$ and $\pm(1.0–0.0)$. The third class is roughly similar to the errors in the $R_2$ measurements and hence may not be considered as significant. Thus it follows that as we move from 8 to 6 M urea the residues 9–12, 16, 55–56, 75, 77, 111, 124, 144, 149, 157 show large propensities for conformational transitions followed by residues 18, 67, 80, and 102. The important numbers among these are: 55–56 at the N terminus, 157 at the C terminus, both of which are cleavage sites of the TFR-PR, and 102,111 at the flaps of the PR domain. As we move to 4 M, mostly the same regions exhibit variations, but the magnitudes are somewhat reduced, except for the stretch 82–84 at the active site, which shows enhancements indicating large conformational transitions. The general reduction may indicate a tendency toward formation of stable contacts. This trend continues as we move to 2 M, where we see a large decrease in the contribution from the conformational transitions at milli- to microsecond time scales. This seems to indicate formation of relatively more rigid contacts. A more detailed characterization would require NOE quantitation and structure calculations, but this is hampered by the tendency of the protein to aggregate and precipitate.

The $^3J(H^N$-$H^\alpha)$ coupling constant, which has the main chain $\phi$ torsion angle dependence, is an NMR parameter that can be rigorously analyzed to get an insight into the secondary structural elements that define the conformational preferences (30). A value in the order of 3–5 Hz corresponds to $\alpha$-helix, 8–11 Hz corresponds to $\beta$-sheet, and 6.0–7.5 Hz, which essentially is an average of the $\alpha$, $\beta$ values corresponds to random coil. It is also observed that the random coil value for any residue is influenced by its N-terminal neighbor and thus two sets of values have been reported for each residue, depending upon whether the N-terminal neighbor belongs to one of the two groups of residues (22) (group I: Phe, Trp, His, Tyr, Ile, Thr, Val, and group II: remaining residues except glycine). Thus, under any given experimental conditions, deviation of the observed coupling constants from the sequence-dependent random coil value, $(J_{obs} - J_{rc})$, which we call as secondary coupling constants analogous to secondary chemical shifts, would throw valuable light on the secondary structural propensities along the polypeptide chain. Negative secondary coupling constants would indicate $\alpha$-helical propensities and positive secondary coupling constants would indicate $\beta$ propensities.

Fig. 8*a* shows the fine structures of the peaks in the HSQC spectra from which the couplings were derived. The measured values range from 5.0 to 9.1 Hz along the sequence in all the cases from 8 to 2 M urea as shown in Fig. 8*b*, and the average

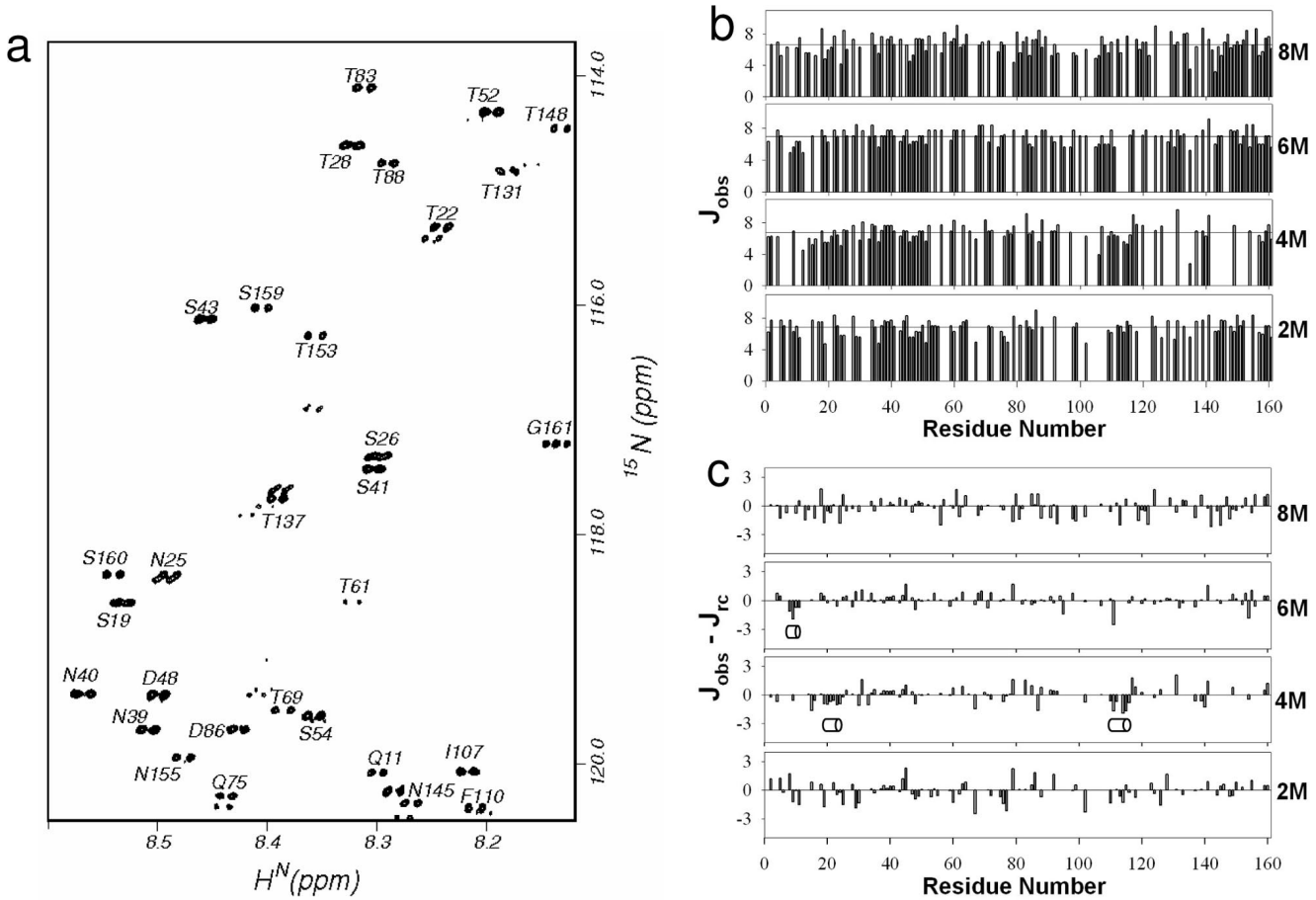


FIG. 8. **Secondary coupling constants.** *a*, portion of the high resolution HSQC spectrum at 6 M urea is shown to illustrate the high resolution of the peaks, which has enabled measurement of $H^N$-$H^\alpha$ coupling constants from peak separations. *b*, measured coupling constants from such spectra at 8, 6, 4, and 2 M urea concentrations. *c*, deviations of the measured coupling constants from the random coil values at the four urea concentrations. Contiguous stretches have been marked with *empty cylinders*.

value (indicated by *horizontal line*) is roughly the same (~6.8 Hz); the estimated error in the measured coupling constants is ~1 Hz. The calculated secondary coupling constants are shown in Fig. 8*c*. As expected, many of the secondary coupling values are zero or close to zero because of random coil characteristics; however there are also few residues that show either positive or negative deviations larger than 1 Hz. Notable among these are the contiguous stretches at 8–12 in 6 M and at 19–25 and 110–116 in 4 M data, which may be taken to indicate some conformational transitions. These have been marked on the figure by empty cylinders and all of them correspond to $\alpha$ propensities. Interestingly, those in the PR region are non-native type.

Fig. 9 summarizes all the observations with regard to the residual structures in 8 M urea, which represent intrinsic initial preferences, and the folding transitions as the urea concentration is reduced. An important inference that can be derived from these at a glance is that folding is not a unidirectional process, in that a structure once formed at any particular intermediate stage does not continue to remain all along till the end. Structure forming-breaking events occur continuously in the progress toward the folded state. It is interesting to note that the TFR region in the precursor, which is considered to be largely unstructured, also has some intrinsic preferences. These may influence the preferences in the PR domain as well by direct interaction. The PR domain has not only native preferences but also many non-native preferences, which must be removed for the protein to fold properly to its native form. In the PR alone this does seem to happen and a fully folded dimeric protein is obtained. However, in the inactive precursor, failure to remove the TFR portion prevents removal of non-native contacts, leads to possible interaction between the TFR and PR regions, which in turn, results in a misfolded state, highly prone to aggregation at the NMR concentrations.

## CONCLUSIONS

We have tried here to obtain residue level insights into the intrinsic conformational preferences in the precursor TFP-$p6^{pol}$-PR-$C^{nn}$ and the characteristics of the folding intermediates by investigating the structural and dynamic features in the species created by systematic variation of urea concentration in the solution. The protein seems to have intrinsic preferences for many non-native contacts in addition to several native preferences, and the presence of TFR prevents proper folding of the protein to the native dimeric structure. However, these non-native contacts seem to get removed after cleavage of the TFR in the normal course. By extrapolation, blocking of this cleavage by an active site mutation seems to prevent removal of the non-native contacts, which in turn leads to misfolding, and consequent aggregation of the protein. Possibly there is an interaction between the TFR region and the PR region, which hinders such a change. Indeed, previous reports have indicated existence of such an interaction (26). An interesting observation in our precursor was that the C-terminal extension led to some retardation of its autoprocessing activity, as shown by the induction SDS-PAGE analysis and the MALDI data. This may suggest an interaction between the C-terminal extension and the PR domain as well. These results viewed in the background

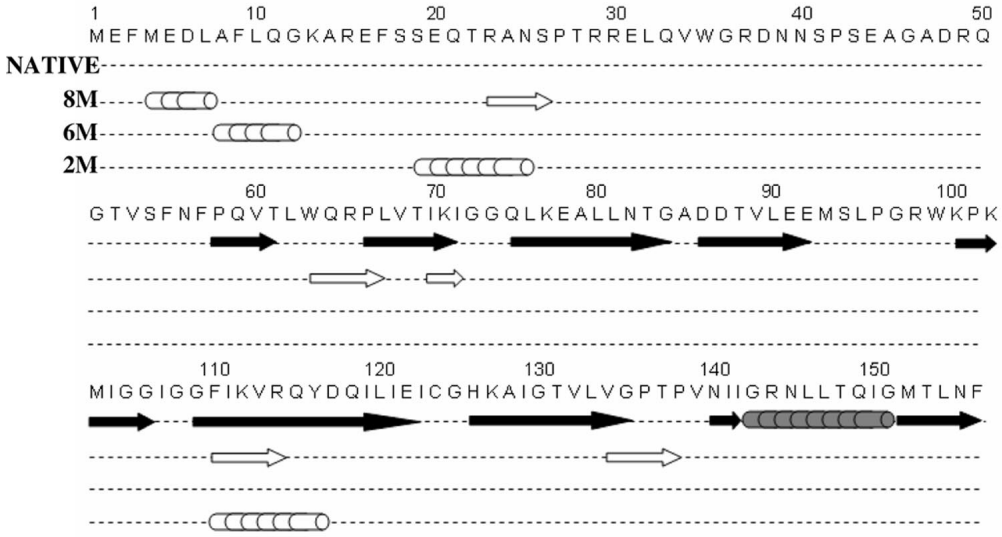


FIG. 9. **Summary of the intrinsic and folding preferences along the sequence of the precursor ($C^{nn}$ stretch is excluded) as the urea concentration is changed.** The native folds in the PR alone are also shown for identification of native and non-native preferences. *Empty arrows* indicate $\beta$-strand propensities and *empty cylinders* indicate $\alpha$-helical propensities. For the 8 M case only those stretches are marked that are common in the $C^{\alpha}$ and $C'$ secondary chemical shifts.

of previous results by Ishima *et al.*, where another precursor, which differed from the one studied here by virtue of several mutations and differences in the TFR sequence at the N terminus, indicates that the sequence plays a major role in dictating the folding propensities of the precursor. In real life, mutations occur frequently in HIV-1 protease precursors because of the poor fidelity of the reverse transcriptase of the virus. As the different mutant proteins would have different folding characteristics, they can also be expected to have different activities and responses to substrates and inhibitors. Taken together, it emerges that folding plays a crucial role in the regulation of precursor processing, and hence in the protease functioning in the viral life cycle. This study, to our knowledge gives the first full-length residue level description of folding characteristics of an HIV-1 protease precursor, and provides useful insights into how folding can have a regulatory role in its autoprocessing and *vice versa*. Our accomplishments here represent another significant advance, namely that this is the longest unfolded polypeptide chain assigned and studied by NMR so far.

*Acknowledgments*—We thank the National Facility for High Field NMR at the Tata Institute of Fundamental research for all the facilities. The clone for the HIV-1 protease-tethered dimer was a kind gift from Dr. M. V. Hosur of Bhabha Atomic Research Center, Mumbai. We thank Geetanjali Dhotre for the MALDI experiments.

REFERENCES

1. Vaishnav, Y. N., and Wong-Staal, F. (1991) *Annu. Rev. Biochem.* **60,** 577–630
2. Skalka, A. M. (1989) *Cell* **56,** 911–913
3. Kay, J., and Dunn, B. M. (1990) *Biochem. BioPhys. Acta* **1048,** 1–18
4. Kohl, N. E., Emini, E. A., Schleif, W. A., Davis, L. J., Heimbach, J. C., Dixon, R. A. F., Scolnick, E. M., and Sigal, I. S. (1988) *Proc. Natl. Acad. Sci. U. S. A.* **85,** 4686–4690
5. Kramer, R. A., Schaber, M. D., Skalka, A. M., Ganguly, K., Wong-Staal, F., and Reddy, E. P. (1986) *Science* **231,** 1580–1584
6. Seelmeier, S., Schmidt, H., Turk, V., and Von der Helm, K. (1988) *Proc. Natl. Acad. Sci. U. S. A.* **85,** 6612–6616
7. Le Grice, S. F. J., Mills, J., and Mous, J. (1988) *EMBO J.* **7,** 2547–2553
8. Darke, P. L., Leu, C. T., Davis, L. J., Heimbach, J. C., Diehl, R. E., Hill, W. S., Dixon, R. A., and Sigal, I. S. (1989) *J. Biol. Chem.* **264,** 2307–2312
9. Wlodawer, A., Miller, M., Jaskolski, M., Sathyanarayana, B. K., Baldwin, E., Weber, I. T., Selk, L. M., Clawson, L., Schneider, J., and Kent, S. B. (1989) *Science* **245,** 616–621
10. Oroszlan, S., and Luftig, R. B. (1990) *Curr. Top. Microbiol. Immunol.* **157,** 153–185
11. Candotti, D., Chappey, C., Rosenheim, M., M'Pele, P., Huraux, J. M., and Agut, H. (1994) *C. R. Acad. Sci. III* **317,** 183–189
12. Schwarzinger, S., Wright, P. E., and Dyson, H. J. (2002) *Biochemistry* **41,** 12681–12686
13. Burstein, H., Bizub, D., Kotler, M., Schatz, G., Vogt, V. M., and Skalka, A. M. (1992) *J. Virol.* **66,** 1781–1785
14. Louis, J. M., McDonald, R. A., Nashed, N. T., Wondrak, E. M., Jerina, D. M., Oroszlan, S., and Mora, P. T. (1991) *Eur. J. Biochem.* **199,** 361–369
15. Kaplan, A. H., Zack, J. A., Knigge, M., Paul, D. A., Kempf, D. J., Norbeck, D. W., and Swanstrom, R. (1993) *J. Virol.* **67,**4050–4055
16. Rose, J. R., Babe, L. M., and Craik, C. S. (1995) *J. Virol.* **69,** 2751–2758
17. Karacostas, V., Wolffe, E. J., Nagashima, K., Gonda, M. A., and Moss, B. (1993) *Virology* **193,** 661–671
18. Krausslich, H. G. (1991) *Proc. Natl. Acad. Sci. U. S. A.* **88,** 3213–3217
19. Wondrak, E. M., Nashed, N. T., Haber, M. T., Jerina, D. M., and Louis, J. M. (1996) *J. Biol. Chem.* **271,** 4477–4481
20. Pettit, S. C, Everitt, L. E, Choudhury, S, Dunn, B. M, and Kaplan, A. H. (2004) *J. Virol.* **78,** 8477–8485
21. Louis, J. M., Nashed, N. T., Parris, K. D., Kimmel, A. R., and Jerina, D. M. (1994) *Proc. Natl. Acad. Sci. U. S. A.* **91,** 7970–7974
22. Penkett, C. J., Redfield, C., Dodd, I., Hubbard, J., McBay, D. L., Mossakowska, D. E., Smith, R. A., Dobson, C. M., and Smith, L. J. (1997) *J. Mol. Biol.* **274,** 152–159
23. Dunn, B. M., Gustchina, A., Wlodawer, A., and Kay, J. (1994) *Methods Enzymol.* **241,** 254–278
24. Beissinger, M., Paulus, C., Bayer, P., Wolf, H., Rosch, P., and Wagner, R. (1996) *Eur. J. Biochem.* **237,** 383–392
25. Louis, J. M., Dyda, F., Nashed, N. T., Kimmel, A. R., and Davies, D. R. (1998) *Biochemistry* **37,** 2105–2110
26. Paulus, C., Hellebrand, S., Tessmer, U., Wolf, H., Krauslich, H. G, and Wagner, R. (1999) *J. Biol. Chem.* **274,** 21539–21543
27. Dautin, N., Karimova, G., and Ladant, D. (2003) *J. Virol.* **77,** 8216–8226
28. Louis, J. M., Wondrak, E. M., Kimmel, A. R., Wingfield, P. T., and Nashed, N. T. (1999) *J. Biol. Chem.* **274,** 23437–23442
29. Louis, M., Clore, G. M., and Gronenborn, A. M. (1999) *Nat. Stuct. Biol.* **6,** 868–875
30. Wüthrich, K. (1986) *NMR of Proteins and Nucleic Acids*, pp. 166–168, Wiley-Interscience, UK
31. Partin, K., Zybarth, G., Ehrlich, L., DeCrombrugghe, M., Wimmer, E., and Carter, C. (1991) *Proc. Natl. Acad. Sci. U. S. A.* **88,** 4776–4780
32. Richards, A. D., Phylip, L. H., Farmerie, W. G., Scarborough, P. E., Alvarez, A., Dunn, B. M., Hirel, P. H., Konvalinka, J., Strop, P., and Pavlickova, L. (1990) *J. Biol. Chem.* **265,** 7733–7736
33. Ishima, R., Torchia, D. A., Shannon, M. L., Angela, M. G., and Louis, J. M. (2003) *J. Biol. Chem.* **278,** 43311–43319
34. Pillai, B., Kannan, K. K., and Hosur, M. V. (2001) *PROTEINS, Struct. Funct. Genet.* **43,** 57–64
35. Panchal, S. C., Pillai, B., Hosur, M. V., and Hosur, R. V. (2000) *Curr. Sci.* **79,** 1684–1695
36. Farrow, N. A., Muhandiram, R., Singer, A. U., Pascal, S. M., Kay, C. M., Gish, G., Shoelson, S. E., Pawson, T., Forman-Kay, J. D., and Kay, L. E. (1994) *Biochemistry* **33,** 5984–6003
37. Louis, J. M., Ishima, R., Nesheiwat, I., Pannell, L. K., Lynch, S. M., Torchia, D. A., and Gronenborn, A. M. (2003) *J. Biol. Chem.* **278,** 6085–6092
38. Ferentz, A. E., and Wagner, G. (2000) *Quart. Rev. Biophys.* **33,** 29–65
39. Panchal, S. C., Bhavesh, N. S., and Hosur, R. V. (2001) *J. Biomol. NMR* **20,** 135–147
40. Bhavesh, N. S., Panchal, S. C., and Hosur, R. V. (2001) *Biochemistry* **40,** 14727–14735
41. Fesik, S., and Zuiderweg, E. R. P. (1990) *Quart. Rev. Biophys.* **23,** 97–131
42. Grzesiek, S., and Bax, A. (1992) *J. Magn. Reson.* **99,** 201–207
43. Grzesiek, S., and Bax, A. (1992) *J. Am. Chem. Soc.* **114,** 6291–6293
44. Juneja, J., and Udgaonkar, J. B. (2003) *Curr. Sci.* **84,** 157–172

45. Shortle D. (1996) *Curr. Opin. Struct. Biol.* **6,** 24–30
46. Dyson, H. J., and Wright, P. E. (2001) *Methods Enzymol.* **339,** 258–270
47. Dyson, H. J., and Wright, P. E. (2002) *Adv. Protein. Chem.* **62,** 311–340
48. Wishart, D. S., and Sykes, B. D. (1994) *Methods Enzymol.* **239,** 363–392
49. Schwarzinger, S., Kroon, G. J. A., Foss, T. R., Wright, P. E., and Dyson, H. J. (2000) *J. Biomol. NMR* **18,** 43–48
50. Schwarzinger, S., Kroon, G. J. A., Foss, T. R., Chung, J., Wright, P. E., and Dyson, H. J. (2001) *J. Am. Chem. Soc.* **123,** 2970–2978
51. Wishart, D. S., Bigam, C. G., Holm, A., Hodges, R. S., and Sykes, B. D. (1995) *J. Biomol. NMR* **5,** 67–81
52. Bhavesh, N. S., Sinha R., Mohan, P. M., and Hosur R. V. (2003) *J. Biol. Chem.* **278,** 19980–19985
53. Kazmirski, S. L., Wong, K. B., Freund, S. M. V., Tan, Y. J., Fersht, A. R., and Daggett, V. (2001) *Proc. Natl. Acad. Sci. U. S. A.* **98,** 4349–4354
54. Klein-Seetharaman, J., Oikawa, M., Grimshaw, S. B., Wirmer, J., Duchardt, E., Ueda, T., Imoto, T., Smith, L. J., Dobson, C. M., and Schwalbe, H. (2002) *Science* **295,** 1719–1722
55. Eliezer, D., Chung, J., Dyson, H. J., and Wright, P. E. (2000) *Biochemistry* **39,** 2894–2901
56. Yao, J., Chung, J., Eliezer, D., Wright, P. E., and Dyson, H. J. (2001) *Biochemistry* **40,** 3561–3571
57. Eliezer, D., Yao, J., Dyson, H. J., and Wright, P. E. (1998) *Nat. Struct. Biol.* **5,** 148–155