DOI: 10.1002/anie.200604087

Enzyme Total Synthesis

Convergent Chemical Synthesis and Crystal Structure of a 203 Amino Acid "Covalent Dimer" HIV-1 Protease Enzyme Molecule**

*Vladimir Yu. Torbeev and Stephen B. H. Kent**

The total chemical synthesis of proteins with sizes larger than about 15 kDa is still a challenging task, even when utilizing modern methods for the ligation of unprotected peptides.[1] The most effective ligation chemistry is the thioester-mediated amide-forming reaction at Cys residues ("native chemical ligation"),[2] and peptides are typically ligated sequentially in the C-to-N terminal direction.[3,4] As a consequence of handling and other losses, synthesis by sequential reactions is inefficient (even in the case of "one-pot" ligations[4]) and consequently the yield of the final polypeptide is low. Recent advances in convergent methodology for the total chemical synthesis of proteins have been proposed to improve the situation.[5]

In our recently reported "kinetically controlled ligation" strategy,[5] the peptide 1-($^{\alpha}$thioarylester) selectively reacts with a Cys-peptide 2-($^{\alpha}$thioalkylester)—in the absence of added thiol—to form the peptide 1-peptide 2-($^{\alpha}$thioalkylester) product in high yield, because of the higher intrinsic reactivity of $^{\alpha}$thioarylesters. This simple concept has resulted in two important implications. First, synthesis (including sequential ligation) from the N-terminal segment towards the C-terminal segment has become possible. Second, two large polypeptides can be assembled in this way, one having a thioester moiety on the C terminus and the other one having a Cys residue on the N terminus. Native chemical ligation of these two large polypeptides at the final stage of the synthesis constitutes a fully convergent approach to the total synthesis of proteins.

We are undertaking detailed studies of the enzymatic mechanism of HIV-1 protease, one of the targets in the therapeutic treatment of AIDS.[6] In its native form, the HIV-1 protease enzyme molecule is a homodimer of two polypeptide chains each containing 99 amino acid residues and with a single active site formed at the dimer interface.[7] The chemical analogues we are constructing to investigate the catalytic mechanism will incorporate different functionalities in the polypeptide chains of the two monomers. To enable nonsymmetric incorporation of functionalities (or labels), the two 99-residue monomers have to be covalently joined through a short linker. Previous approaches to covalent linking have included recombinant expression of polypeptides of approximately 210 residues,[8] or have employed a synthetic strategy involving the directed formation of a disulfide bond between the two chains.[9] Although enzymes made in this way have led to insights about the catalytic mechanism,[9b] the overall synthesis is inefficient and, thus, a more robust synthetic route was required for further work. Herein we report the convergent chemical synthesis of a polypeptide chain with 203 amino acids[10] from four peptide segments. We demonstrate the full catalytic activity of the resulting enzyme molecule and report its high-resolution X-ray structure.

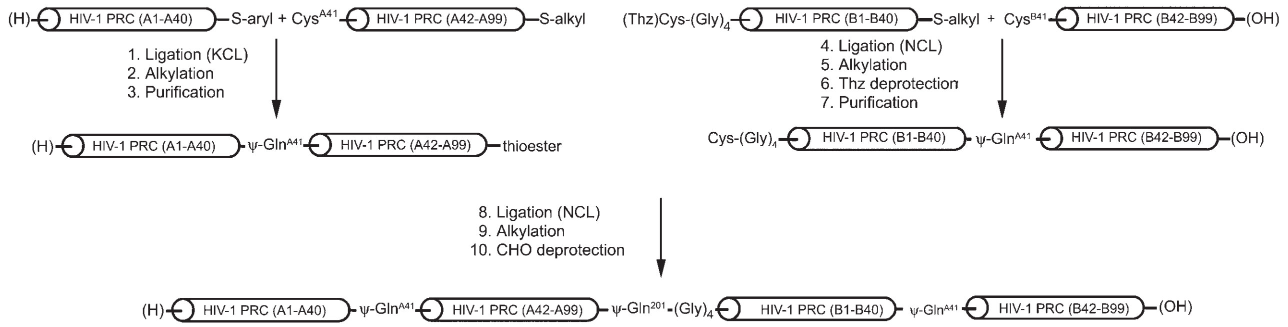
The first step in the convergent synthesis of the target polypeptide (Scheme 1) is the kinetically controlled ligation of the two peptide segments (A1-A40)-($^{\alpha}$thioarylester) (**1**) and Cys-(A42-A99)-($^{\alpha}$thioalkylester) (**2**). Segment **1** was obtained after transthioesterification of (A1-A40)-($^{\alpha}$thioalkylester) with an excess of 4-mercaptophenylacetic acid.[11] Ligation was performed at pH 6.3 to slow down all the reactions and thus get better overall control. Two main byproducts were present in the reaction mixture (Figure 1, Scheme 2). One is the branched thioester **7**, formed by reaction of the ligation product (A1-Cys^{A41}-A99)-($^{\alpha}$thioalkylester) (**3**) with **1**. The second is the internal thiolactone **8**, formed from intramolecular transthioesterification of the ligation product. After an empirically determined optimal reaction time of one hour, excess 4-mercaptophenyl acetic acid was added to give a total concentration of 200 mM at pH 6.0; this leads to breakdown of the branched thioester **7**, thereby releasing more of the ligation product **3** and regenerating starting peptide **1**, which can further ligate with any remaining **2**. Moreover, both the internal thiolactone **8** and the ligation product **3** undergo transthioesterification to form the desired ligation product **4**. The sulfhydryl functionality of Cys^{A41} was subsequently capped with 2-bromoacetamide to form ψ-Gln^{A41} at the ligation site.

The segment Cys-Gly$_4$-(B1-B99) was synthesized by conventional native chemical ligation.[2] Two peptides Thz-Gly$_4$-(B1-B40)-($^{\alpha}$thioalkylester) and Cys-(B42-B99) were ligated at pH 7.0 using 4-mercaptophenylacetic acid as a catalyst.[11] Residue Cys^{B41} at the ligation site was then alkylated with 2-bromoacetamide and the ligation product

[*] V. Yu. Torbeev, Prof. Dr. S. B. H. Kent
Department of Chemistry
Institute for Biophysical Dynamics
Gordon Center for Integrative Science
The University of Chicago
929 East 57th Street, Chicago, IL 60637 (USA)
Fax: (+1) 773-702-0439
E-mail: skent@uchicago.edu

[**] We acknowledge funding from the DOE "Genomics to Life" program (grant no. DE-FG02-04ER63786). Use of the Advanced Photon Source (Argonne, IL) was supported by the U.S. Department of Energy, Basic Energy Sciences, Office of Science, under contract no. W-31-109-Eng-38. Use of the GM/CA-CAT beamline has been funded in whole or in part with Federal funds from the National Cancer Instutute (Y1-CO-1020) and the National Institute of General Medical Sciences (Y1-GM-1104). We acknowledge the help and advice from Dr. Valentina A. Terechko (University of Chicago) on X-ray structure refinement.

Supporting information for this article is available on the WWW under http://www.angewandte.org or from the author.


(H)-HIV-1 PRC (A1-A40)-S-aryl + Cys^A41-HIV-1 PRC (A42-A99)-S-alkyl
1. Ligation (KCL)
2. Alkylation
3. Purification
(H)-HIV-1 PRC (A1-A40)-ψ-Gln^A41-HIV-1 PRC (A42-A99)-thioester
(Thz)Cys-(Gly)4-HIV-1 PRC (B1-B40)-S-alkyl + Cys^B41-HIV-1 PRC (B42-B99)-(OH)
4. Ligation (NCL)
5. Alkylation
6. Thz deprotection
7. Purification
Cys-(Gly)4-HIV-1 PRC (B1-B40)-ψ-Gln^A41-HIV-1 PRC (B42-B99)-(OH)
8. Ligation (NCL)
9. Alkylation
10. CHO deprotection
(H)-HIV-1 PRC (A1-A40)-ψ-Gln^A41-HIV-1 PRC (A42-A99)-ψ-Gln^201-(Gly)4-HIV-1 PRC (B1-B40)-ψ-Gln^A41-HIV-1 PRC (B42-B99)-(OH)

Scheme 1. Convergent synthesis of the "covalent dimer" HIV-1 protease. Designations S-aryl = 4-mercaptophenylacetic acid thioester, S-alkyl = 3-mercaptopropionic acid tetraarginine amide, KCL = kinetically controlled ligation, NCL = native chemical ligation, Thz = thiazolidine. For the sequence see Ref. [10], and for detailed experimental procedures see the Experimental Section and the Supporting Information.

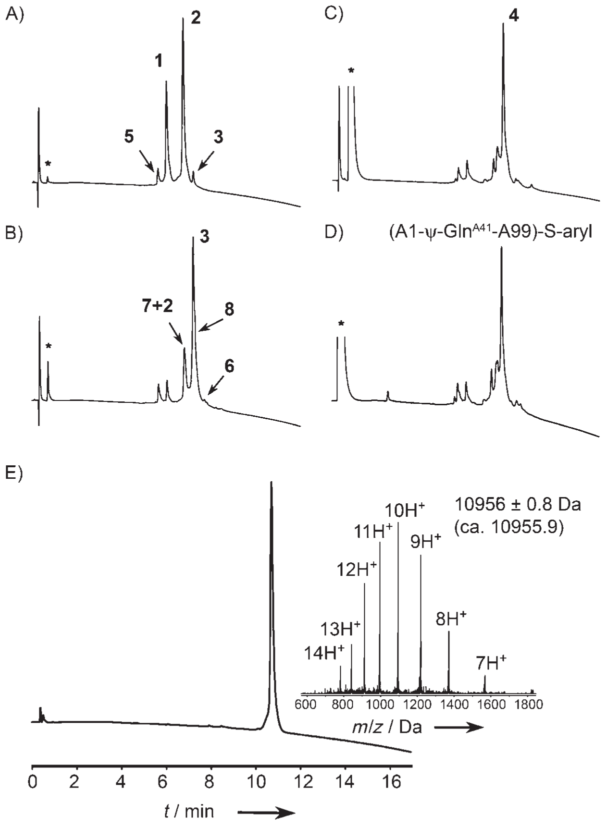
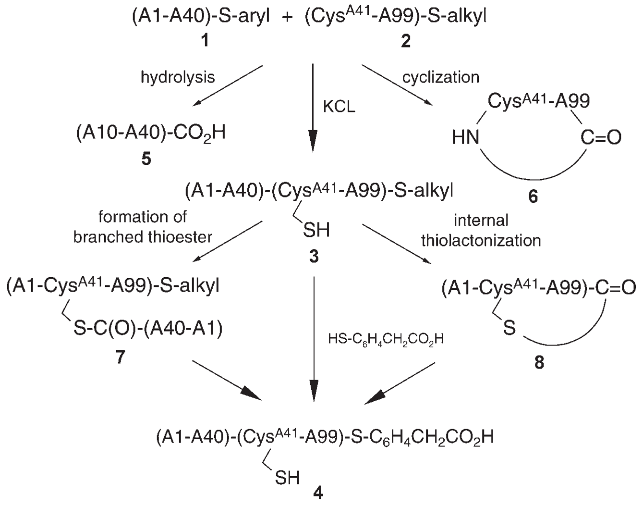


Figure 1. Analytical HPLC traces ($\lambda = 214$ nm) of kinetically controlled ligation of (A1-A40)-$^{\alpha}COSC_6H_4CH_2COOH$ (**1**) and Cys-(A42-A99)-$^{\alpha}COSCH_2CH_2Arg_4$ (**2**). A) $t < 1$ min, pH 5.0; **5** indicates the product (A1-A40)-CO_2H arising from hydrolysis (see Scheme 2), **3** is the product (A1-A99)-$^{\alpha}COSCH_2CH_2Arg_4$ arising from ligation. B) $t = 1$ h, pH 6.3; **7** + **2** indicates coeluting (A1-A99)-(A1-A40)-$^{\alpha}COSCH_2CH_2Arg_4$ and recovered (A41-A99)-$^{\alpha}COSCH_2CH_2Arg_4$, **6** is cyclic (Cys^{A41}-A99), and **8** (right shoulder) is the internal thiolactone. C) After addition of 200 mM 4-mercaptophenylacetic acid at pH 6.0, for 3 h. D) After S-alkylation with 2-bromoacetamide. E) After purification by HPLC. The asterisks indicate 4-mercaptophenylacetic acid (in A–C) or S-alkylated 4-mercaptophenylacetic acid and buffer components (in D).



Scheme 2. Kinetically controlled ligation (KCL) of peptide-($^{\alpha}$thioarylester) **1** and peptide-($^{\alpha}$thioalkylester) **2**. The desired ligation peptide **3** and by-products **7** and **8** were transformed to a single product **4** by treatment with 4-mercaptophenylacetic acid.

treated with $MeONH_2{\cdot}HCl$ to convert the N-terminal thiazolidine into a Cys residue.

The purified segments (A1-A99)-($^{\alpha}$thioarylester) and Cys-Gly$_4$-(B1-B99) were then joined together by native chemical ligation to form a final polypeptide chain consisting of 203 amino acids (Figure 2). The Cys residue at the final ligation site was converted into ψ-Gln by treatment with 2-bromoacetamide. After removal of the formyl protecting groups from the tryptophan residues,[12] the product was purified by reversed-phase HPLC (RP-HPLC; 6.7% overall yield of isolated product based on the limiting peptide segment). The 203-residue synthetic polypeptide was characterized by LC-MS, and further analyzed by Fourier-transform ion cyclotron resonance mass spectrometry (FT-ICR-MS) (Figure 3). Within the limits of experimental certainty, the product had the expected mass (found 21869.8 ± 0.4 Da; calcd 21869.8 Da, average isotope composition).

The synthetic polypeptide was folded by two-step dialysis against acetate buffer at pH 5.6 (29% yield). A standard fluorogenic assay of the enzymatic activity was performed in 50 mM NaOAc and 0.2 M NaCl at pH 5.6 and 37 °C with Abz-Thr-Ile-Nle-Phe(p-NO_2)-Gln-Arg.amide (Abz = 2-aminobenzoyl).[13] The k_{cat} and K_m values of 10.3 ± 0.2 s^{-1} and 27 ±

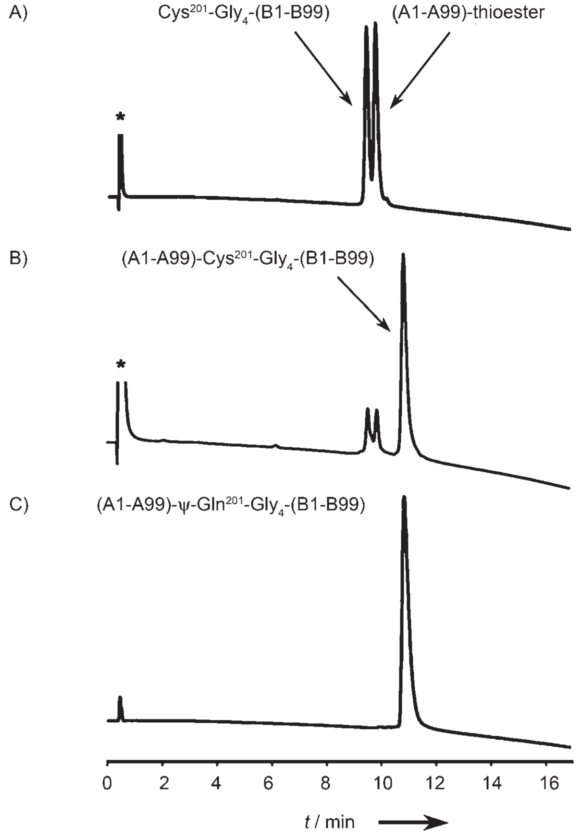


Figure 2. Analytical HPLC traces ($\lambda = 214$ nm) corresponding to the final step in the convergent synthesis of the HIV-1 "covalent dimer" construct. A) $t < 1$ min. B) $t = 5$ h, pH 7.0. C) Product after S-alkylation with 2-bromoacetamide, removal of the formyl protecting groups, and HPLC purification. The asterisks indicate guanidine·HCl (Gn·HCl) and tris(2-carboxyethyl)phosphine (TCEP) in the case of (A) and added 50 mM 4-mercaptophenylacetic acid in (B).

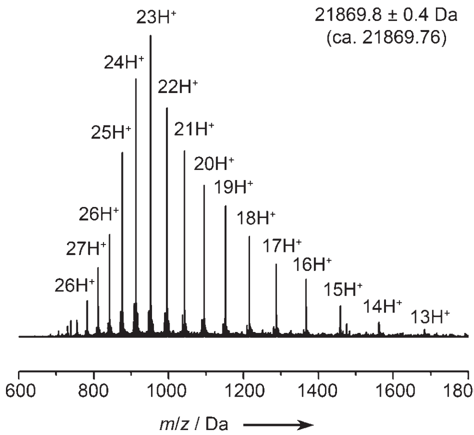


Figure 3. FT-ICR ESI MS spectrum of the "covalent dimer" HIV-1 protease (see the Supporting Information for more details).

1.4 µM, respectively (see the Supporting Information), are in agreement with previously reported data recorded under similar assay conditions.[14] As a control, a chemically synthesized homodimeric HIV-1 protease (that is, 2 × 99 residues) was assayed under the same conditions ($k_{cat} = 9.8 \pm 0.2\ s^{-1}$, $K_m = 25 \pm 1.4$ µM).

X-ray structural analysis was performed to verify if the synthetic protein had the correct three-dimensional fold of the HIV-1 protease covalent dimer enzyme. Crystals grown in the presence of the inhibitor MVT-101 (Ac-Thr-Ile-Nle-ψ-(CH_2NH)-Nle-Gln-Arg.amide) were isomorphous to those of previously reported synthetic and recombinantly expressed HIV-1 proteases, and diffracted to a resolution of 1.65 Å. The X-ray structure of the protein molecule with 203 amino acids (Figure 4) was found to be essentially identical to the previously reported structures of homodimeric HIV-1 protease,[15] as well as to those of recombinantly expressed tethered dimers of HIV-1 protease,[16] with the linker region being partially disordered.

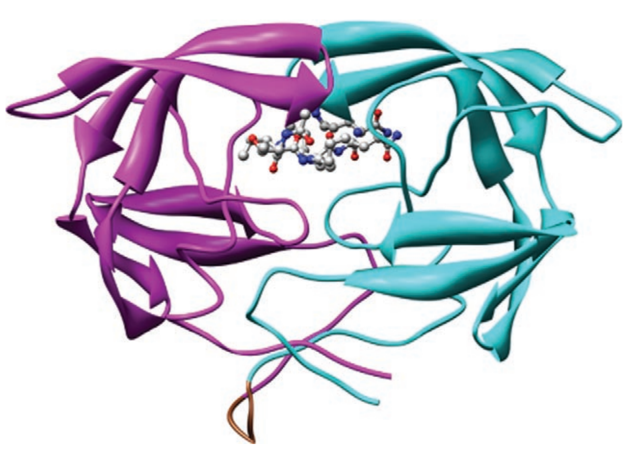


Figure 4. X-ray crystallographic structure of the "covalent dimer" HIV-1 protease complexed with the MVT-101 inhibitor (the linker region ψ-Gln^{201}GlyGlyGlyGly^{205} is shown in red).

This 21870 Da protein with full enzymatic activity and correct three-dimensional structure is, to the best of our knowledge, the largest linear polypeptide chain prepared to date by chemical synthesis. The total synthesis of a protein of this size, in a straightforward fashion, demonstrates the great potential of recently developed methods for the fully convergent chemical synthesis of proteins. Facile synthetic access to the 203-residue "covalent dimer" HIV-1 protease will enable the preparation of a wide range of unique chemical analogues to systematically dissect the molecular basis of the function of this important enzyme.

Experimental Section

In a kinetically controlled ligation, (A1-A40)-$^{\alpha}COSC_6H_4CH_2COOH$ (8.2 mg, 1.8 µmol) and (Cys^{A41}-A99)-$^{\alpha}COSCH_2CH_2Arg_4$ (13.5 mg, 1.9 µmol) were dissolved in aqueous buffer (1.46 mL) containing 6 M Gn·HCl, 0.2 M Na_2HPO_4, and 19 mM TCEP at pH 6.3. After 1 h 4-mercaptophenylacetic acid was added to give a total concentration of about 200 mM and the pH value was adjusted to 6.0. After 3 h, 2-bromoacetamide (52 mg, 0.377 mmol) was added and the pH value adjusted to 6.7. After 15 min, 4-mercaptophenylacetic acid (51 mg, 0.304 mmol) was added to neutralize the excess of 2-bromoacetamide. The product was purified by RP-HPLC with a shallow gradient of water/acetonitrile with 0.1 % trifluoroacetic acid (TFA). LC-MS: found: 10956 ± 0.8 Da, calcd: 10955.9 Da (Figure 1). Yield of isolated product 5.7 mg (0.52 μmol, 29 %). For the synthesis of Cys-Gly$_4$-(B1-B99) see the Supporting Information.

In the final native chemical ligation, (A1-A99)-$^{\alpha}$COSC$_6$H$_4$CH$_2$COOH (4.8 mg, 0.44 μmol) and Cys-Gly$_4$-(B1-B99) (5.5 mg, 0.49 μmol) were dissolved in buffer (1.6mL) containing 8 M Gn·HCl, 0.1 M Na$_2$HPO$_4$, and 25 mM TCEP. 4-mercaptophenylacetic acid was added to give a concentration of 50 mM and the pH value was adjusted to 7.0. After 12 h, the reaction mixture was diluted with buffer (1 mL), and 2-bromoacetamide (100 mg, 0.72 mmol) was added at pH 6.7. After 15 min, the reaction was quenched with an excess of 4-mercaptophenylacetic acid. Deformylation was performed by treatment with a mixture of 2-mercaptoethanol and piperidine (1:1 (v/v), 3.6 mL) on ice for 15 min, and then neutralizing with HCl. The reaction mixture was diluted twofold with buffer (6 M Gn·HCl, 0.1 M Na$_2$HPO$_4$) and purified by RP-HPLC. LC-MS: found: 21 869.8 ± 0.4 Da, calcd: 21 869.8 Da. Yield of isolated product 2.2 mg (0.1 μmol, 23 %). The overall yield based on the limiting peptide segment is 6.7 %.

For more experimental details see the Supporting Information.

Received: October 4, 2006
Revised: December 4, 2006
Published online: January 24, 2007

Keywords: convergent synthesis · enzymes · native chemical ligation · peptides · protein synthesis

[1] P. E. Dawson, S. B. H. Kent, *Annu. Rev. Biochem.* **2000**, *69*, 923–960.

[2] P. E. Dawson, T. W. Muir, I. Clark-Lewis, S. B. H. Kent, *Science* **1994**, *266*, 776–779.

[3] a) G. G. Kochendoerfer et al., *Science* **2003**, *299*, 884–887, see the Supporting Information; b) T. M. Hackeng, J. M. Griffin, P. E. Dawson, *Proc. Natl. Acad. Sci. USA* **1999**, *96*, 10068–10073; c) Z. Wu, J. Alexandratos, B. Ericksen, J. Lubkowski, R. C. Gallo, W. Lu, *Proc. Natl. Acad. Sci. USA* **2004**, *101*, 11587–11592.

[4] D. Bang, S. B. H. Kent, *Angew. Chem.* **2004**, *116*, 2588–2592; *Angew. Chem. Int. Ed.* **2004**, *43*, 2534–2538.

[5] D. Bang, B. L. Pentelute, S. B. H. Kent, *Angew. Chem.* **2006**, *118*, 4089–4092; *Angew. Chem. Int. Ed.* **2006**, *45*, 3985–3988.

[6] R. J. Pomerzantz, D. L. Horn, *Nat. Med.* **2003**, *9*, 867–873.

[7] a) J. Schneider, S. B. H. Kent, *Cell* **1988**, *54*, 363–368; b) A. Wlodawer, M. Miller, M. Jaskolski, B. K. Sathyanarayana, E. Baldwin, I. T. Weber, L. M. Selk, L. Clawson, J. Schneider, S. B. H. Kent, *Science* **1989**, *245*, 616–621.

[8] a) Y.-S. E. Cheng, F. H. Yin, S. Foundling, D. Blomstrom, C. A. Kettner, *Proc. Natl. Acad. Sci. USA* **1990**, *87*, 9660–9664; b) C. L. DiIanni, L. J. Davis, M. K. Holloway, W. K. Herber, P. L. Darke, N. E. Kohl, R. A. F. Dixon, *J. Biol. Chem.* **1990**, *265*, 17348–17354; c) D. Bizub, I. T. Weber, C. E. Cameron, J. P. Leis, A. M. Skalka, *J. Biol. Chem.* **1991**, *266*, 4951–4958; d) H.-G. Kräusslich, *Proc. Natl. Acad. Sci. USA* **1991**, *88*, 3213–3217; e) J. T. Griffiths, L. A. Tomchak, J. S. Mills, M. C. Graves, N. D. Cook, B. M. Dunn, J. Kay, *J. Biol. Chem.* **1994**, *269*, 4787–4793.

[9] a) M. Baca, T. W. Muir, M. Schnölzer, S. B. H. Kent, *J. Am. Chem. Soc.* **1995**, *117*, 1881–1887; b) M. Baca, S. B. H. Kent, *Tetrahedron* **2000**, *56*, 9503–9513.

[10] Sequence of the tethered construct of HIV-1 protease (from the N to the C terminus): PQITLWKRPL^{A10} VTIRIGGQLK^{A20} EALLDTGADD^{A30} TVIEE*Nle*NLPG^{A40} *ψ-Gln*WKPK*Nle*IG-GI^{A50} GGFIKVRQYD^{A60} QIPVEI*Abu*GHK^{A70} AIGTVLVGPT^{A80} PVNIIGRNLL^{A90} TQIG*Abu*TLNF^{A99} *ψ-Gln*^{201}GGGG^{205} PQITLWKRPL^{B10} VTIRIGGQLK^{B20} EALLDTGADD^{B30} TVIEE*Nle*NLPG^{B40} *ψ-Gln*WKPK*Nle*IG-GI^{B50} GGFIKVRQYD^{B60} QIPVEI*Abu*GHK^{B70} AIGTVLVGPT^{B80} PVNIIGRNLL^{B90} TQIG*Abu*TLNF^{B99}. Unnatural amino acids are shown in italics in three-letter code. *Nle* = norleucine, *Abu* = α-aminobutyric acid, *ψ-Gln* = pseudo-homoglutamine. Residues from the 99-residue section at the N terminus (part A) are specified by the letter "A" placed before the number of the residue, correspondingly the 99-residue part at the C terminus (part B) has the letter "B" before the number. The five amino acid linker region is numbered from 201 to 205. Ligation sites are underlined.

[11] E. C. B. Johnson, S. B. H. Kent, *J. Am. Chem. Soc.* **2006**, *128*, 7140–7141.

[12] Four tryptophan residues are present: W^{A6}, W^{A42}, W^{B6}, and W^{B42}.

[13] M. V. Toth, G. R. Marshall, *Int. J. Pept. Protein Res.* **1990**, *36*, 544–550.

[14] a) T. D. Meek, E. J. Rodriguez, T. S. Angeles, *Methods Enzymol.* **1994**, *241*, 127–156; b) Z. Q. Beck, L. Harvio, P. E. Dawson, J. H. Elder, E. L. Madison, *Virology* **2000**, *274*, 391–401.

[15] a) A. Wlodawer, J. W. Erickson, *Annu. Rev. Biochem.* **1993**, *62*, 543–585; b) A. Wlodawer, J. Vondrasek, *Annu. Rev. Biophys. Biomol. Struct.* **1998**, *27*, 249–284.

[16] a) T. N. Bhat, E. T. Baldwin, B. Liu, Y.-S. E. Cheng, J. W. Erickson, *Nat. Struct. Biol.* **1994**, *1*, 552–556; b) B. Pillai, K. K. Kannan, M. V. Hosur, *Proteins Struct. Funct. Genet.* **2001**, *43*, 57–64.