(2). In all these cases, the regions of native β sheet that are detected first are those involved in tertiary interactions with a helix (or helices) that are also stably formed early in the folding process. In the native structure of the all–β sheet protein IL-1β, on the other hand, the three pseudosymmetric elements of β sheet do not form such interactions. Each β strand contributes two or three nonpolar residues to a hydrophobic core that depends for its stability on the hydrophobic and van der Waals interactions of a large number of tightly packed side chains. We therefore suggest that folding to the stable native structure for this type of protein involves the rapid formation of β structure around a nonpolar core, followed by the much slower stabilization of native secondary structure that accompanies the progressive final tight packing of the core groups or those groups external to the β sheets, or both.

REFERENCES AND NOTES

1. H. Roder, G. A. Elöve, S. W. Englander, *Nature* **335**, 700 (1988).
2. J. B. Udgaonkar and R. L. Baldwin, *ibid.*, p. 694; M. Bycroft, A. Matouschek, J. T. Kellis, A. R. Fersht, *ibid.* **346**, 488 (1990); M. S. Briggs and H. Roder, *Proc. Natl. Acad. Sci. U.S.A.* **89**, 2017 (1992); J. Lu and F. W. Dahlquist, *Biochemistry* **31**, 4749 (1992); S. E. Radford, C. M. Dobson, P. A. Evans, *Nature* **358**, 302 (1992).
3. A. V. Finkelstein, *Proteins* **9**, 23 (1991).
4. K. A. Dill, *Biochemistry* **24**, 1501 (1985).
5. IL-1β is made up of 12 antiparallel β strands connected by turns, short loops, or long loops (*7*). The long loop (residues 30 to 39) connecting β strands 3 and 4 is unusual in that it contains a type II turn (residues 32 to 35) followed by 2.3 turns of a 3_{10} helix (residues 33 to 39) characterized by four CO(*i*)-NH(*i* + 3) hydrogen bonds; the NH protons of residues 36 to 39, however, are not slowly exchanging. Because this loop is surface-exposed and the short 3_{10} helix lies outside the core of the protein, IL-1β can still be considered to be an all–β sheet protein.
6. P. C. Driscoll, G. M. Clore, D. Marion, P. T. Wingfield, A. M. Gronenborn, *Biochemistry* **29**, 3542 (1990); G. M. Clore, A. Bax, P. C. Driscoll, P. T. Wingfield, A. M. Gronenborn, *ibid.*, p. 8172.
7. G. M. Clore, P. T. Wingfield, A. M. Gronenborn, *ibid.* **30**, 2315 (1991).
8. B. C. Finzel *et al.*, *J. Mol. Biol.* **209**, 779 (1989); J. P. Priestle, H. P. Schär, M. G. Grütter, *Proc. Natl. Acad. Sci. U.S.A.* **86**, 9667 (1989); B. Veerpandian *et al.*, *Proteins* **12**, 10 (1992).
9. B. Shaanan *et al.*, *Science* **257**, 961 (1992).
10. S. Craig, U. Schmeissner, P. T. Wingfield, R. H. Pain, *Biochemistry* **26**, 3570 (1987).
11. O. B. Ptitsyn, R. H. Pain, G. V. Semisotnov, E. Zerovnik, O. I. Razgulyaev, *FEBS Lett.* **262**, 20 (1990).
12. Rapid kinetics of the change in far-UV CD were measured with a Biologic SFM-3 stopped-flow mixing system fitted with a cell of path length 1 mm installed on a Jobin Yvon Dichrographe Mk 6 (*14*). Folding at 4°C was initiated by mixing of 1 volume of unfolded IL-1β (1.5 mg/ml) in 2 M GuHCl with 5 volumes of 100 mM sodium acetate (pH 6.5). Typically, the results of 12 separate runs were accumulated. At the end of each experiment the cell was washed with acetate buffer. Controls were performed in which unfolded protein in GuHCl buffer was mixed with the same buffer.
13. Folding at 4°C was initiated by mixing of 1 volume of unfolded IL-1β (1 mg/ml) in 2 M GuHCl with 5 volumes of 10 mM MES, 100 mM NaCl, and, in the case of the ANS experiments, 0.1 mM ANS (pH 6.5) with a Hi-Tech stopped-flow apparatus. Tryptophan and ANS fluorescence were excited at 290 and 380 nm, respectively, and measured above 400 and 430 nm, respectively, with high-pass filters.
14. P. Varley *et al.*, unpublished results.
15. Previous experiments in which the protein was unfolded in GuHCl for different lengths of time before the initiation of refolding have indicated that this slow phase is not due to cis-trans proline isomerization giving rise to both fast and slow refolding populations of unfolded and partially folded molecules (*10*).
16. Quenched-flow mixing was performed with a Biologic QFM-5 module. IL-1β (~4 mg/ml and uniformly labeled to >95% with ^{15}N), prepared as described in (*7*), was unfolded in D_2O for at least 1 hour in 3 M guanidinium deuterochloride (GuDCl)–50 mM sodium d_3-acetate (pH 5.05) to allow complete denaturation and H-D exchange. The slightly lower pH than that used for the fluorescence and CD kinetics was necessitated by the greater tendency of IL-1β to aggregate at the higher concentrations required for the D-H exchange experiment. Refolding at 4°C was initiated by dilution with 7 volumes of folding buffer [50 mM sodium d_3-acetate (pH 5.05) in H_2O for folding times of ≤1 s and in D_2O for folding times >1 s]. After a selected time T of refolding, the sample was pulsed to rapidly exchange deuterons for protons on nonprotected amide groups by dilution with 5 volumes of pulse buffer [100 mM glycine-NaOH (pH 9.75)] to produce a final pH of 9.3. After 16 ms, the pulse was ended by a further 1:1 dilution with 0.5 M sodium d_3-acetate (pH 5.05) to give a final pH of 5.10, and the protein was allowed to fold to completion (as determined by NMR). For T = 25 s, mixing was carried out manually, and exchange of nonprotected amide groups, 25 s after the initiation of folding, was attained by dilution of the folding mixture with 10 volumes of 500 mM sodium acetate in H_2O (pH 5.05). The resulting samples were then concentrated and exchanged into 50 mM sodium d_3-acetate in D_2O (pH 5.05) with Amicon Centriprep 10 concentrators. Refolding was carried out at relatively low protein concentrations to avoid complications arising from aggregation. The extent of protection was then measured by recording a ^1H-^{15}N Overbodenhausen correlation spectrum [A. Bax *et al.*, *J. Magn. Reson.* **86**, 304 (1990)] on a Bruker AMX 600 spectrometer. The concentration of IL-1β in the NMR samples was 0.3 to 0.4 mM. To compare cross-peak intensities between samples taken at different refolding times T, we normalized the spectra (relative to the T = 0 s spectrum) to correct for the different concentrations of protein in the samples (as measured spectrophotometrically by the absorbance at 280 nm and by the intensity of the resonances in the aliphatic region of the ^1H-NMR spectrum) and the concentration of D_2O relative to H_2O at the time of the exchange pulse.
17. P. C. Driscoll, A. M. Gronenborn, P. T. Wingfield, G. M. Clore, *Biochemistry* **29**, 4668 (1990).
18. G. M. Clore, A. Bax, P. T. Wingfield, A. M. Gronenborn, *ibid.*, p. 5671.
19. K. Kuwajima, H. Yamaya, S. Miwa, S. Sugai, T. Nagamura, *FEBS Lett.* **221**, 115 (1987).
20. A. F. Chaffotte, Y. Guillou, M. E. Goldberg, *Biochemistry* **31**, 9694 (1992).
21. H-D exchange experiments on native IL-1β at pH 9.3, using ^1H-^{15}N Overbodenhausen correlation spectroscopy to monitor the NH → ND exchange as a function of time after taking up a lyophilized sample of IL-1β in D_2O, indicated that after 1 hour there was no change in the intensities of the 47 slowly exchanging NH protons (that is, those with a $t_{1/2}$ > 3000 min at pH 5).
22. P. J. Kraulis, *J. Appl. Crystallogr.* **24**, 946 (1991).

23. Supported by the Intramural AIDS Targeted Antiviral Program of the Office of the Director of the National Institutes of Health (to G.M.C. and A.M.G.) and by a grant from the Science and Engineering Research Council (to R.H.P.).

7 December 1992; accepted 12 February 1993

Substrate Phage: Selection of Protease Substrates by Monovalent Phage Display

David J. Matthews and James A. Wells*

A method is described here for identifying good protease substrates among approximately 10^7 possible sequences. A library of fusion proteins was constructed containing an amino-terminal domain used to bind to an affinity support, followed by a randomized protease substrate sequence and the carboxyl-terminal domain of M13 gene III. Each fusion protein was displayed as a single copy on filamentous phagemid particles (substrate phage). Phage were then bound to an affinity support and treated with the protease of interest. Phage with good protease substrates were released, whereas phage with substrates that resisted proteolysis remained bound. After several rounds of binding, proteolysis, and phagemid propagation, sensitive and resistant substrate sequences were identified for two different proteases, a variant of subtilisin and factor X_a. The technique may also be useful for studying the sequence specificity of a variety of posttranslational modifications.

Proteolysis is a common form of posttranslational modification and is important in regulation and protein turnover (*1*). Knowledge of protease specificity aids in the identification of biologically relevant substrates, helps direct the design of specific inhibitors, and is useful in applying proteases for site-specific proteolysis (*2*). Substrate sequences for proteases often extend over seven or eight contiguous residues (*3*), and thus one protease can potentially interact with a vast number of possible substrates.

We present a method, called substrate phage, in which more than 10^7 potential substrates can be tested concurrently. Each substrate sequence is displayed as a single

Department of Protein Engineering, Genentech, South San Francisco, CA 94080.

*To whom correspondence should be addressed.

copy on phagemid particles between an affinity domain and a truncated form of the gene III protein of M13 (Fig. 1). The affinity domain is a variant of human growth hormone (hGH) (*4*) that binds tightly to the hGH-binding protein (hGHbp) (*5*). Substrate phage are bound to immobilized hGHbp and then cleaved by incubation with a protease. The protease cleaves phage containing good substrate sequences in the randomized region; phage containing noncleaved sequences remain bound to the immobilized support and can be eluted with a low pH buffer. The protease-sensitive or protease-resistant phage pools are propagated by infection of *Escherichia coli*, and the process is repeated to further enrich for phage harboring desirable substrate or nonsubstrate sequences.

We used the substrate phage method to investigate the substrate specificity of a variant of the bacterial serine protease subtilisin BPN′, in which the catalytic His^{64} was replaced with Ala (H64A) (*6*, *7*). The mutated enzyme acquired specificity for cleaving substrates containing His in the P2 position (*8*) by a substrate-assisted catalytic mechanism. Further mutation of this enzyme produced a hexamutant (S24C-H64A-E156S-G166A-G169A-Y217L, referred to here as the H64A subtilisin variant) that was more active and useful for cleaving fusion proteins at specific sites containing histidines (*9*). However, some P2 His–containing sequences were much better substrates than others, and evidence indicated that one substrate could be cleaved with the His assisting from the P1′ position (Fig. 2). Thus, we desired to determine the best substrates for P2 His–directed cleavage or possible P1′ His–directed cleavage because this information is critical for the efficient use of this enzyme to cleave NH_2-terminal (P2 His) or COOH-terminal (P1′ His) fusion proteins.

We inserted a known P2 His–containing substrate sequence (AAHYTRQ) between the hGH variant and a truncated form of gene III protein (*10*) and bound it to hGHbp immobilized on microtiter plates. After treatment with the H64A subtilisin variant, ~100 times more phagemid particles harboring the substrate sequence insert were released compared to the number of similar phagemid particles released that lack the insert (*11*).

We constructed two substrate phage libraries by inserting the sequence GPG-$G(X)_5$GGPG or GPAA$(X)_5$AAPG between the hGH variant and gene III, where X represents any of the 20 naturally occurring amino acids (*7*). Thus, each library represents 3.2×10^6 possible protein sequences. We flanked the random substrate sequence with Gly-Pro followed by Pro-Gly to break any secondary structure imposed by the hGH or gene III domains, and we included Gly-Gly linkers in one library to possibly improve protease susceptibility by increasing segmental flexibility (*12*). We allowed each library of substrate phage (containing 5×10^6 or 2×10^6 independent transformants for the Gly-Gly–flanked and Ala-Ala–flanked libraries, respectively) to bind to the hGHbp in wells of polystyrene plates. The wells were treated with protease, and the released phage were propagated (protease-sensitive pool). Those that remained bound were eluted at pH 2 and propagated (protease-resistant pool).

After three rounds of protease selection from the GPGG$(X)_5$GGPG library, all clones that were protease-sensitive contained a His residue in their sequence (Table 1). His occurred almost exclusively at position 2 or 4 within the randomized sequence. After compensating for the predicted codon frequency for each amino acid type, we found His, Met, and Tyr to be overrepresented in the protease-sensitive library. Closer analysis of the sequences in Table 1 shows a consensus sequence with His preferentially flanked by Tyr, Met, Leu, and Thr residues. Hydrophobic residues, particularly Tyr, Met, Leu, and Phe, are preferred at the P1 position of subtilisin (*13*). No Pro or Cys residues were found among substrate sequences. Sequences containing a Pro from positions P1 to P2′ are poor substrates (*9*).

Clones that were resistant to three rounds of protease selection exhibited no sequence consensus. Only one sequence containing His was isolated (GGHPSEPGG). However, this sequence cannot be cleaved with His assisting from the P2 position because subtilisin is unable to cleave substrates with Pro at P1. Thus, protease-sensitive sequences resemble good substrates, whereas resistant clones resemble poor substrates. We also sorted the substrate phage library containing the GP-AA$(X)_5$AAPG linker between hGH and the gene III with results that were similar to those seen in Table 1 (*11*). After six rounds of selection, all selected clones contained at least one His residue in position two or four within the randomized region. None of the resistant clones contained a His residue, and many contained one or more Pro residues.

Optimal substrates are those with the highest ratios of the turnover number (k_{cat}) [for the enzyme (E)-substrate (S) complex (E·S)] to the Michaelis constant (K_m). For the enzyme and substrate concentrations used during the protease selection, we expect that substrate phage are sorted on the basis of relative k_{cat}/K_m values (*14*). We developed a simple assay to evaluate the selected substrate sequences in the context of a fusion protein containing the hGH and substrate sequences linked to alkaline phosphatase (AP). Seven individual clones from

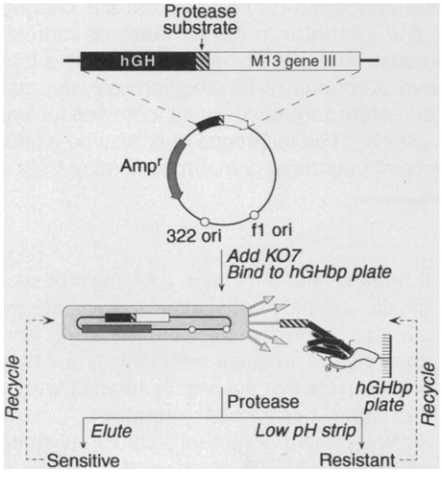

Fig. 1. Scheme for protease substrate phage selection.

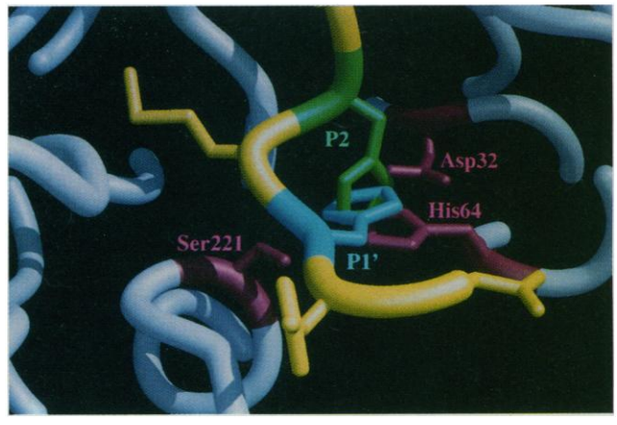

Fig. 2. Computer model showing the proposed interactions of subtilisin BPN′ with substrates containing His at the P2 (green) or P1′ (light blue) position. The substrate is colored yellow, the main chain of the enzyme is white, and the catalytic triad (Ser^{221}, His^{64}, and Asp^{32}) is magenta. Molecular modeling studies were based on x-ray crystal structures of subtilisin BPN′ [Brookhaven protein data bank reference 1SBT (*27*)] and Streptomyces subtilisin inhibitor [Brookhaven protein data bank reference 2SSI (*28*)]. The substrate His was introduced at the P2 or P1′ position, and the torsion angles were varied so that the Nδ1 and Nε2 nitrogens of the substrate were in close proximity to those of His^{64}. In both cases, no unreasonable torsion angles were introduced in the substrate. Thus, a His from either the P2 or P1′ position of the substrate can virtually occupy the position of the missing catalytic His^{64} in the H64A subtilisin variant.

the protease-selected library (Table 1) were inserted into an hGH-AP expression vector (*15*). Fusion proteins were expressed in *E. coli* KS330 and assayed for release of AP activity after incubation with the H64A subtilisin variant. Dimeric hGH-AP fusion proteins were immobilized on beads coated with hGHbp and incubated with the H64A subtilisin variant to release AP activity (*16*). The assay used identical conditions to those in the phage selection procedure (*16*). The NH_2-termini of the AP cleavage products were also determined (*17*).

The relative amounts of cleavage for the seven different sequences in the AP fusion assay varied more than tenfold between the different substrate sequences (Table 2). No activity was observed above the background concentration for the nonsubstrate sequence HPSEP. In a high proportion of sequences (four out of seven), cleavage was directed from the P1′ residue. Two of the most efficiently cleaved sequences isolated (TSMHT and YHLKM) were cleaved with P1′ His–assisted catalysis. Thus, catalysis assisted by P1′ His can be as efficient as that assisted by P2 His. In each case, one mode of cleavage was significantly preferred, even though some of the substrates (such as DGYHY) were expected to be cleaved from either the P2 or P1′ His positions. The P1′ His–assisted cleavage is interesting because once the acyl-enzyme intermediate is formed the P1′ His residue may possibly be released and therefore not be available to assist in deacylation. This is consistent with other studies that suggest that P2 His is critical in acylation but not in deacylation (*20*).

Fusion proteins are a convenient means of expressing a recombinant protein that is linked to an affinity handle to facilitate purification (*2*). Knowledge of good P2 His substrates can be used to design efficient sites to cleave the affinity handle linked on the NH_2-terminus of the protein of interest. Conversely, the information about efficient P1′ His substrates may allow the use of the H64A subtilisin variant for cleaving affinity handles linked to the COOH-terminus of a protein.

To investigate the generality of the substrate phage technique, we used the library containing the GPAA$(X)_5$AAPG linker to study the specificity of human factor X_a, a blood-clotting protease that has also been used for cleaving fusion proteins (*2*, *19*). Factor X_a has specificity determinants that extend over more than four contiguous residues and is known to have a strong preference for cleaving P1 Arg substrates (*20*, *21*). After four rounds of protease selection, almost every clone contained at least one Arg in the substrate linker sequence, and four of them contained two Arg residues (Table 3). Of the 16 sequences isolated, 6 match tripeptide *p*-nitroanilide substrates (*20*) or tripeptide-4-methyl coumaryl-7-amide substrates (*21*) that are efficiently hydrolyzed by factor X_a, including EGR, LGR (isolated twice), EAR,

Table 1. Sequences sensitive and resistant to cleavage by the H64A variant of subtilisin BPN′ that were isolated from the GPGG$(X)_5$GGPG library. Substrate libraries were constructed by cassette mutagenesis as described (*4*). The phagemid phGH-LIB-g3 (*24*) was digested with Apa I and then with Sal I, precipitated with ethanol, and purified by gel electrophoresis. A substrate cassette was synthesized comprising the oligonucleotides 5′-CGGTGGTNNSNNSNN-SNNSNNSGGTGGTCCTGGC-3′ and 5′-TC-GAGCCAGGACCACCSNNSNNSNNSNNSN-NACCACCGGGCC-3′, where N represents an equal mixture of G, A, T, and C and S represents an equal mixture of G and C bases. This creates a peptide sequence GPGG$(X)_5$GGPG between the Apa I and Sal I sites of the cassette. We annealed the cassette by heating the oligonucleotides to 90°C and then slowly cooling them to room temperature. The resulting oligonucleotide cassette was ligated with approximately 1 μg of the digested vector (in a 1:10 molar ratio of vector to cassette), and the ligation product was transformed into *E. coli* XL-1 Blue by electroporation as described (*25*). A control transformation was also done with no oligonucleotide cassette present. We used transformed cells to prepare a phage library by growing them overnight with M13K07 helper phage that were added at a concentration to give a multiplicity of infection (MOI) of approximately 100. Sorting of the phage library was done essentially as described (*10*) but in 1-ml polystyrene plates. Protease-sensitive phage were eluted after incubation for 10 min at room temperature with 500 nM H64A subtilisin variant (*9*). The plates were washed in PBS containing 0.01% Tween 20. Any phage still bound to the plates were eluted with 50 mM glycine (pH 2.0). (This phage pool contained protease-resistant phage.) We propagated protease-sensitive and protease-resistant phage by infecting a fresh culture of *E. coli* XL-1 and growing the bacteria overnight with M13K07 (MOI of 100). After transduction of 1 hour at 37°C, the cells were centrifuged and resuspended in 2YT [1 ml; 16 g of bacto-tryptone, 10 g of bacto-yeast extract, and 5 g of NaCl per liter (pH 7.0)]. This selection procedure was then repeated. After three rounds of selection, the ratio of phage eluted with enzyme to nonspecifically eluted phage was ~40:1. At this stage, 18 clones from the protease-sensitive and protease-resistant libraries were sequenced (*26*). Amino acids are denoted by single-letter codes (*7*); O denotes an amber stop codon, which is translated as Gln in the *supE* strain of *E. coli*, XL-1.

Sensitive	Resistant
FHMNV	SPAQN
NHYTL	LSPNM
THYFL	MPRTF
KHAYL	KSMVA
YHMMA	INDTL
YHLKM	DVNKP
THTTQ	ARRTV
KHYTI	GNSQS
QHYVN	EWALL
DGYHL	ISPLI
TSNHI	ALMDS
LLRHT	TNFSA
ATLHL	TGNNT
AOMHM	SRISL
TSMHT	NLELN
YSLHV	VYSTN
YAMHF	HPSEP
FRAMH	PSKSY

Table 2. Results of hGH-AP fusion assay of clones from the GPGG$(X)_5$GGPG substrate phage library. We assayed substrates by constructing hGH-AP fusion proteins. The amount of AP activity released and the site of cleavage (indicated by ↓) were determined as described (*16*, *17*). The numbers have been corrected for the background release of AP in the absence of protease (50 to 90 ng/ml). For the nonsubstrate sequence HPSEP, no activity above the background concentration was detected.

Sequence	AP released (ng/ml)	His position
NHY ↓ TL	401 ± 22	P2
TSM ↓ HT	221 ± 18	P1′
Y ↓ HLKM	89 ± 13	P1′
FHM ↓ NV	56 ± 14	P2
DGY ↓ HY	47 ± 14	P1′
THY ↓ FL	35 ± 13	P2
TSN ↓ HI	24 ± 11	P1′
HPSEP	0	

Table 3. Sequences sensitive or resistant to cleavage by human factor X_a that were selected from the GPAA$(X)_5$AAPG library after four rounds. The library was constructed in a fashion analogous to that for the GPGG$(X)_5$GGPG library described in Table 1, except that the oligonucleotide directed the production of two Ala codons flanking the random pentacodon sequence. Production of phagemid particles, binding to plates, and protease selection were done as described in Table 1, except that the concentration of human factor X_a (Haematologic Technologies, Essex Junction, Vermont) was 50 nM in the first round and 5 nM in subsequent rounds. Protease treatment was done in 50 mM tris (pH 7.5), 150 mM NaCl, 0.005% Triton X-100, and 2 mM $CaCl_2$ for 15 min at 25°C.

Sensitive	Resistant
LLGRT	ISNSN
EGRGR	FTLDM
NRGEG	YTVKP
NFHPR	STLLG
QMVLR	TFQVR
SLLGR	AVLHV
RSLTR	SLPHY
QYRFR	RSKEN
NKYFR	KMSFY
TREAR	AIINK
TPSTR	LLLTH
HSRFR	PTPLP
RIADA	PKYEA
PRTRA	IQYRH
QGAQN	
WSRKV	

STR, and LTR. Also, several sequences were isolated with Phe in the putative P2 position; such sequences are known to be relatively good factor X_a substrates (*20*). Although at this point we do not know the cleavage sites and relative rates of hydrolysis for these selected sequences, the k_{cat}/K_m values for EGR and LGR tripeptides (the only two compared directly) are within a factor of 2 to 4 of each other (*20, 21*). Substrate phage sequences containing Arg were found far more frequently than sequences containing Lys, and P1 Arg substrates are preferred over P1 Lys substrates by up to a factor of $\sim 10^4$ (*20*). In contrast, only a few Arg-containing sequences were found in the protease-resistant library. Thus, as with the H64A subtilisin variant, many of the sequences selected for hydrolysis by factor X_a resemble good substrates.

We have also conducted some experiments with the aspartyl protease from human immunodeficiency virus type–1 (HIV-1) (*22*). We inserted a known HIV-1 protease substrate (SQNYPIVQ) between hGH and the truncated form of gene III protein. Approximately ten times more phagemid particles were released, as compared with similar phagemid particles that lacked the substrate sequence (*11*).

A number of factors may limit the generality of the substrate phage method. First, the preferential site of cleavage must occur at the random substrate linker and not at the affinity domain or gene III protein. In the event that proteolysis of competing sites is too rapid to allow selection, mutations can be made or a different affinity domain can be chosen that resists proteolysis. It is unlikely that the protease will digest the phage itself because it is resistant to proteolysis. A second limitation is that proteases in *E. coli* may cleave some of the substrate sequences. Thus, it may be desirable to use protease-deficient strains of *E. coli* so that the substrate phage cannot be cleaved by endogenous proteases. Editing during expression can be detected by a lack of particular residues in both the sensitive and resistant substrate phage pools. For example, the fact that Cys was missing from both resistant and sensitive substrate phage pools in our experiments suggests that it was edited out. This may reflect difficulties in expressing a properly folded in hGH–gene III fusion protein if it were to contain an unpaired Cys in the substrate linker. Finally, proteases with broad specificities will probably not yield a consensus substrate sequence by substrate phage selection. However, this is a limitation of any method used to define protease specificity.

There are some 300 different amino acid derivatives that occur in nature (*23*). Substrate phage are potentially useful for determining the substrate specificity for post-translational modifications. The method requires only that a randomized substrate domain be displayed in such a way that it can be specifically modified in vitro and separated from unmodified substrates. For example, to study protein phosphorylation one could use an antibody to phosphotyrosine for separating substrate phage that react with a tyrosine kinase. We view the substrate phage method as a screening tool to be used to provide a manageable number of substrates from a myriad of possibilities that can then be further evaluated by traditional kinetic methods.

REFERENCES AND NOTES

1. H. Neurath, *Trends Biochem. Sci.* **14**, 268 (1989).
2. M. Uhlen and T. Moks, *Methods Enzymol.* **185**, 124 (1990); P. Carter, in *Protein Purification: From Molecular Mechanisms to Large-Scale Processes*, M. R. Ladisch, R. C. Willson, C.-D. Painton, S. E. Builder, Eds. (ACS Symposium Series 427, American Chemical Society, Washington, DC, 1990), chap. 13, p. 181; B. Nilsson *et al.*, *Curr. Opin. Struct. Biol.* **2**, 569 (1992).
3. J. Kraut, *Annu. Rev. Biochem.* **46**, 331 (1977).
4. H. B. Lowman, S. H. Bass, N. Simpson, J. A. Wells, *Biochemistry* **30**, 10832 (1991).
5. G. Fuh *et al.*, *J. Biol. Chem.* **265**, 3111 (1990).
6. P. Carter and J. A. Wells, *Science* **237**, 394 (1987).
7. Mutations are designated by the wild-type residue, followed by the residue's position and then the mutant residue in single-letter codes. Multiple mutants are indicated by a series of single mutants separated by a slash. Amino acids are denoted by single-letter codes (Ala, A; Arg, R; Asn, N; Asp, D; Cys, C; Glu, E; Gln, Q; Gly, G; His, H; Ile, I; Leu, L; Lys, K; Met, M; Phe, F; Pro, P; Ser, S; Thr, T; Trp, W; Tyr, Y; and Val, V).
8. We used the notation of Schechter and Berger [I. Schechter and A. Berger, *Biochem. Biophys. Res. Commun.* **27**, 157 (1967)] to describe substrate residues with the form NH_2-P*n* . . .-P2-P1-P1′-P2′ . . . P*n*-COOH, where hydrolysis occurs between the P1 and P1′ residues.
9. P. Carter *et al.*, *Proteins Struct. Funct. Genet.* **6**, 240 (1989).
10. We constructed phagemids in which a tight-binding variant of hGH (*4*) was linked directly to a truncated form of M13 gene III protein (pDM0390) or in which a substrate sequence (AAHYTRQ) was inserted between hGH and the truncated gene III protein (pDM0411). First, a partial Pvu II digest was done on a derivative of phGHam-g3 containing the gene for a variant of hGH with the mutations R167N-D171N-K172S-F176Y-I178T-F10A-M14W-H18D-H21N (*4*). We then isolated the 747-bp Pvu II fragment from the partial digestion and ligated it into the large Pvu II fragment of phGH-M13gIII [S. Bass, R. Greene, J. A. Wells, *Proteins Struct. Funct. Genet.* **8**, 309 (1990)] to create the vector pDM0390. We performed site-directed mutagenesis of pDM0390 with the oligonucleotide 5′TTCGGGCCCTTCGCTGCTCACTATACGCGTCAGTCGACTGACCTGCCT-3′ to introduce a subtilisin cleavage site (AAHYTRQ) between hGH and bacteriophage gene III to create the vector pDM0411. Wells of a microtiter plate were coated with hGHbp (2 μg/ml) in 50 mM $NaHCO_3$ buffer (pH 9.8) overnight and blocked for 2 hours with phosphate-buffered saline (PBS; 10 mM sodium phosphate at pH 7.4 and 150 mM NaCl) containing 0.5% (w/v) bovine serum albumin (BSA). Equal numbers of phage derived from the phagemid constructs pDM0390 and pDM0411 [in 10 mM tris-HCl (pH 7.6), 1 mM EDTA, and 100 mM NaCl] were incubated in the wells at room temperature for 2 hours with gentle agitation. The plate was washed thoroughly (four times with PBS plus 0.01% Tween 20, interspersed with 10-min incubations in wash buffer). The plate was briefly equilibrated in 20 mM tris (pH 8.6) with 100 mM NaCl. The H64A subtilisin variant (*9*) (500 nM final concentration) was added, and we incubated the wells for 10 min with gentle agitation to specifically elute the phage. A control elution with no enzyme was also performed. Phage titers [in colony-forming units (CFUs)] were measured for each sample.
11. D. J. Matthews and J. A. Wells, unpublished results.
12. A. Fontana *et al.*, *Biochemistry* **25**, 1847 (1986).
13. D. A. Estell *et al.*, *Science* **233**, 659 (1986).
14. At enzyme and substrate concentrations well below the K_m for the enzyme, the relative rates of hydrolysis reflect differences in k_{cat}/K_m ratios [A. R. Fersht, *Enzyme Structure and Mechanism* (Freeman, New York, ed. 3, 1985), pp. 105–106]. The K_m value for typical subtilisin substrates is ~100 μM (*13*), which is much higher than the total concentration of phage bound to the wells (~1 pM; 10^8 phagemid particles in 10^{-4} liters). However, in order to effect good enrichments of the substrate phage over nonsubstrate phage, we found it was preferable to use relatively high concentrations of subtilisin (up to 0.5 μM). At these low substrate concentrations where enzyme is in excess, the rate of hydrolysis (v) is given by
$$v = \frac{k_{cat}(E_o)(S_o)}{K_m + E_o}$$
where E_o is the initial concentration of enzyme, S_o is the initial substrate concentration, and K_m is the Michaelis constant. When $K_m >> E_o$, this equation simplifies to
$$v = \frac{k_{cat}(E_o)(S_o)}{K_m}$$
For amide bond hydrolysis, acylation is rate limiting, and therefore K_s (the dissociation constant for the E·S complex) ≅ K_m [J. A. Wells, B. C. Cunningham, T. P. Graycar, D. A. Estell, *Philos. Trans. R. Soc. London Ser. A* **317**, 415 (1986)]. Under these conditions, the relative rates of hydrolysis are proportional to k_{cat}/K_m, and substrate phage selection sorts for the optimal substrate. (However, if $E_o >> K_m$, then the equation simplifies to $v = k_{cat}(S_o)$, and sorting occurs for the substrates on the basis of their relative k_{cat} values).
15. Fusion proteins were prepared with a derivative of the vector pZAP (*9*) that produces a fusion protein in which a synthetic (Z) domain of *Staphylococcus aureus* protein A is fused to *E. coli* AP. The region encoding the protein A signal sequence was replaced by a sequence encoding the heat-stable enterotoxin (STII) signal sequence. Also, a Sal I restriction endonuclease site was engineered into the linker region downstream of the sequence encoding the enzyme cleavage site. We used two oligonucleotide primers, 5′-TGTCACGGCCGAGACTTATAGTCGC-3′ and 5′-CACCGCCAGTCGACCCAGGACCACC-3′, to amplify the hGH and substrate gene sequence from the library vector using the polymerase chain reaction (PCR). We cleaved PCR-amplified hGH-substrate fragment with Xba I and Sal I and cloned it into the pZAP vector to create phGH-AP, an expression vector that makes hGH fused to AP through the substrate library linker region.
16. The phGH-AP vectors were transformed into *E. coli* KS330 cells [K. L. Strauch and J. Beckwith, *Proc. Natl. Acad. Sci. U.S.A.* **85**, 1576 (1988)] and grown overnight in Luria broth at 37°C in the presence of carbenicillin (50 μg/ml) and kanamycin (50 μg/ml). On the next day, we transferred 0.5 ml of culture to low-phosphate medium (25 ml) [C. N. Chang *et al.*, *Gene* **55**, 189 (1987); modified as follows: 0.15% glucose, 0.22% casein hydrolysate, 0.03% yeast extract, 1.6 mM $MgSO_4$, 20 mM NH_4Cl, 50 mM KCl, 20 mM NaCl, and 120 mM triethanolamine (pH 7.4)] containing carbenicillin (25 μg/ml) and kanamycin (25 μg/ml). After incubation with shaking at 30°C for 16 to 20 hours, the cells were centrifuged and frozen at −20°C for 48 hours. We isolated periplasmic proteins by thawing the frozen cell pellets in 0.5 ml of 10 mM tris

(pH 7.6), 1 mM EDTA, and 1 mM phenylmethylsulfonyl fluoride. After centrifugation, the supernatant was added to an equal volume of 90% (w/v) ammonium sulfate and left on ice for 30 min. The samples were centrifuged, and the protein pellets were resuspended in 0.5 ml of 10 mM tris (pH 7.6) with 1 mM EDTA. The relative concentration of each fusion protein was determined by SDS–polyacrylamide gel electrophoresis (SDS-PAGE) under reducing conditions with Coomassie blue staining and laser-scanning densitometry; all sample volumes were adjusted to give the same concentration of protein. The hGHbp was coupled to Affi-Gel 15 (5 ml), (Bio-Rad) according to the manufacturer's instructions and made up to a total volume of 10 ml in 10 mM tris (pH 7.6), and 1 mM EDTA. The affinity matrix (40 μl) was incubated with 100 μl (~1 μg) of hGH-AP fusion protein for 2 hours with slow tumbling at room temperature. The supernatant was removed, and the beads were washed twice in 1 ml of 10 mM tris (pH 7.6) with 1 mM EDTA and once in 0.5 ml of 20 mM tris (pH 8.6) with 100 mM NaCl. The H64A subtilisin variant (100 μl of 500 nM) was added, and the beads were incubated with tumbling for 10 min at room temperature. The supernatant was removed and assayed for AP activity by incubating a sample (20 μl) with 180 μl of 6.6 mM *p*-nitrophenyl phosphate (Sigma) in 0.6 M tris (pH 8.2) for 2 hours at room temperature with shaking and measuring the absorbance change at 405 nm caused by production of *p*-nitroaniline. The amount of AP released with no protease added was also measured.

17. For NH_2-terminal sequence analysis, AP cleavage products were isolated essentially as described (*16*), but 100 μl of Affi-Gel slurry and 200 μl of fusion protein were used for each sample. Also, the protease concentration was increased to 2 μM in a volume of 50 μl, and the cleavage reaction proceeded for 1 hour. The resulting AP digestion products were electroblotted onto polyvinylidene difluoride membrane [P. Matsudaira, *J. Biol. Chem.* **262**, 10035 (1987)], and the NH_2-termini were sequenced with a protein sequencer (Applied Biosystems 473 or 477).
18. P. Carter, L. Abrahmsen, J. A. Wells, *Biochemistry* **30**, 6142 (1991). Cleavage of the FAHF-*p*-nitroanilide substrate by either the wild-type or the H64A variant does not produce a burst of production of *p*-nitroaniline, which indicates that acylation is rate limiting. Furthermore, the k_{cat} for cleavage of a FAQF-thiobenzyl ester substrate (where deacylation is presumed to be rate limiting) by the H64A variant is 1×10^{-2} s^{-1}, whereas the k_{cat} for cleavage of an FAHF-*p*-nitroanilide substrate (where His-assisted acylation is rate limiting) is 2×10^{-2} s^{-1}. This suggests that His-assisted catalysis is critical for acylation but not as important for deacylation.
19. K. Nagai and H. C. Thogersen, *Nature* **309**, 810 (1984).
20. K. Cho *et al.*, *Biochemistry* **23**, 644 (1984); R. Lottenberg, U. Christensen; C. M. Jackson, L. Coleman, *Methods Enzymol.* **80**, 341 (1981).
21. S. Kawabata *et al.*, *Eur. J. Biochem.* **172**, 17 (1988).
22. C. Debouck, *AIDS Res. Human Retroviruses* **8**, 153 (1992).
23. S. C. B. Yan, B. W. Grinnell, F. Wold, *Trends Biochem. Sci.* **14**, 264 (1989).
24. The phagemid phGH-LIB-g3 was derived from phGHam-g3 (*4*). Site-directed mutagenesis of phGHam-g3 with the oligonucleotide 5′-AGCTGTGGCTTCGGGCCCGCCGCCGCGTCGACTGGCGGTGGCTCT-3′ resulted in the amber stop codon being replaced by a short Apa I–Sal I cassette. The wild-type hGH gene was replaced with the gene for a tight-binding hGH molecule by exchange of an Apa I–Hinc II fragment from the phagemid pDM0390 (*10*). The short Apa I–Sal I cassette was excised and replaced by a longer cassette constructed from the oligonucleotides 5′-CTAATGATAATGAATTAATTGATAATGATAATG-3′ and 5′-TCGACATTATCATTATCAATTAATTCATTATCATTAGGGCC-3′ so that the Sal I site was no longer at the end of a fragment and could be cut more efficiently. The cassette contains eight stop codons and introduces a frame-shift mutation between hGH and gene III that ensures that when library cassettes are introduced only plasmids containing a randomized substrate insert will be propagated as hGH-displaying phage.
25. W. J. Dower, J. F. Miller, C. W. Ragsdale, *Nucleic Acids Res.* **16**, 6127 (1988).
26. F. Sanger, S. Nicklen, A. R. Coulson, *Proc. Natl. Acad. Sci. U.S.A.* **74**, 5463 (1977).
27. C. S. Wright, R. A. Alden, J. Kraut, *Nature* **221**, 235 (1969).
28. Y. Mitsui, Y. Satow, Y. Watanabe, Y. Iitaka, *J. Mol. Biol.* **131**, 697 (1979).
29. We thank S. Bass for supplying hGH–gene III and some preliminary substrate phagemid vectors, W. Henzel and S. Wong for help with protein sequencing, the oligonucleotide group at Genentech for supplying synthetic DNA, R. Lazarus and J. Seymour for providing factor X_a and for advice on its use, C. Craik for supplying HIV protease, W. Anstine for graphics, and M. Zoller for comments and critical reading of the manuscript.

28 December 1992; accepted 9 March 1993

Structure of the Retinoid X Receptor α DNA Binding Domain: A Helix Required for Homodimeric DNA Binding

Min S. Lee,* Steven A. Kliewer, Joan Provencal, Peter E. Wright,† Ronald M. Evans†

The three-dimensional solution structure of the DNA binding domain (DBD) of the retinoid X receptor α (RXRα) was determined by nuclear magnetic resonance spectroscopy. The two zinc fingers of the RXR DBD fold to form a single structural domain that consists of two perpendicularly oriented helices and that resembles the corresponding regions of the glucocorticoid and estrogen receptors (GR and ER, respectively). However, in contrast to the DBDs of the GR and ER, the RXR DBD contains an additional helix immediately after the second zinc finger. This third helix mediates both protein-protein and protein-DNA interactions required for cooperative, dimeric binding of the RXR DBD to DNA. Identification of the third helix in the RXR DBD thus defines a structural feature required for selective dimerization of the RXR on hormone response elements composed of half-sites (5′-AGGTCA-3′) arranged as tandem repeats.

The mechanisms by which transcription factors bind to regulatory sequences and control expression of target genes is a central problem in eukaryotic molecular biology. Members of the nuclear hormone receptor superfamily contain a highly conserved region of ~70 amino acids, including two zinc fingers, that is required for specific binding to DNA sequences termed hormone response elements (HREs) (*1*). Typically, members of the family bind as dimers to HREs composed of two copies of a six-nucleotide motif, termed half-sites. A subset of the nuclear receptors, including the GR and ER, bind as homodimers to HRE half-sites oriented as inverted repeats (*1*). In contrast, other members of the nuclear receptor family, including the peroxisome proliferator–activated receptor (PPAR), vitamin D receptor (VDR), thyroid hormone receptor (TR), and retinoic acid receptor (RAR), preferentially bind and activate through HREs composed of half-sites arranged as direct repeats (DRs), with specificity conferred by both the half-site sequence and the number of nucleotides separating the two half-sites (*2*). Instead of binding as homodimers, these receptors form heterodimers with the RXR that bind with high affinity to target DNA (*3*, *4*). In addition to its role in heterodimeric complexes, the RXR also forms a homodimer that activates in response to 9-*cis* retinoic acid through HREs composed of DRs (*5*, *6*).

To determine the structural features of the RXR that promote binding to tandem repeat HREs, we expressed a 94-residue peptide (Fig. 1) that comprised the DBD of RXRα in *Escherichia coli* and purified it to near homogeneity (*7*). In gel mobility-shift assays (*8*), the RXR DBD peptide bound weakly to an oligonucleotide containing a single AGGTCA half-site (Fig. 2A). In contrast, the RXR DBD bound cooperatively to an oligonucleotide that contained two half-sites oriented as direct repeats, which indicates the presence of a dimerization signal in the DBD (Fig. 2A). The isolated RXR DBD retained the binding specificity of the full-length RXR protein (*6*), binding preferentially as a homodimer to a direct repeat of AGGTCA with a single nucleotide spacer relative to the other spacing options (Fig. 2B).

M. S. Lee and P. E. Wright, Department of Molecular Biology, The Scripps Research Institute, La Jolla, CA 92037.
S. A. Kliewer, J. Provencal, R. M. Evans, Howard Hughes Medical Institute, The Salk Institute for Biological Studies, La Jolla, CA 92037.

*Present address: Molecumetics Institute, 2023 120th Avenue, N.E., Suite 400, Bellevue, WA 98005.
†To whom correspondence should be addressed.