Probing the chemical basis of binding activity in an SH3 domain by protein signature analysis

Tom W Muir*, Philip E Dawson, Michael C Fitzgerald and Stephen BH Kent

Background: Modifying the covalent structure of a protein is an effective empirical route to probing three-dimensional structure and biological function. Here we describe a combinatorial protein chemistry strategy for studying structure–activity relationships in proteins. Our approach (termed 'protein signature analysis') involves functional selection from an array of self-encoded protein analogs prepared by total synthesis, coupled to a simple chemical readout that unambiguously identifies the modified proteins in the resulting active and inactive populations.

Results: Protein signature analysis was used to study the interaction of the amino-terminal SH3 domain from the cellular adaptor protein c-Crk with its cognate proline-rich peptide, C3G. Using a functional selection assay, the qualitative effects of scanning a series of synthetic analog units through the amino-acid sequence of the SH3 domain were evaluated. The analog units were designed to alter both amino-acid sidechains and the polypeptide backbone within the protein. These chemical studies revealed that the sidechain of Asp150 in the SH3 domain is essential for ligand binding and that changes in the structure of the polypeptide backbone can also result in loss of binding activity.

Conclusions: These chemical studies have provided new insight into how ligand binding is related to the covalent structure of the SH3 domain. Protein signature analysis is a powerful and conceptually novel way of studying the molecular and chemical basis of protein function; it combines the advantages of systematic modification of a protein's chemical structure with the practical convenience of combinatorial synthesis.

Address: The Scripps Research Institute, 10550 North Torrey Pines Road, La Jolla, California 92037, USA.

*Present address: Synthetic Protein Chemistry Laboratory, Box 223, The Rockefeller University, 1230 York Avenue, New York, NY 10021, USA.

Correspondence: Tom Muir
e-mail: muirt@rockvax.rockefeller.edu

Key words: combinatorial protein chemistry, protein array, protein signature analysis

Received: **Jul 22 1996**
Revisions requested: **Aug 12 1996**
Revisions received: **Aug 19 1996**
Accepted: **Aug 29 1996**

Chemistry & Biology October 1996, **3**:817–825

Introduction

Understanding how the chemical structure of a protein determines its functional properties is an important current challenge in biological research. Altering the covalent structure of a protein has proven to be an extremely effective way of exploring the molecular basis of biological function. Site-directed mutagenesis of recombinantly expressed proteins offers a powerful route to studying the individual contributions of each amino acid sidechain in a protein to the function of the molecule [1]. Classical mutagenesis techniques are constrained, in a chemical sense, to the covalent structures and stereochemistry of natural (i.e., genetically encoded) amino-acid residues, and to the natural peptidic linkage between them. These restrictions have been loosened by the development of novel *in vitro* expression approaches, which allow a range of unnatural amino acids to be incorporated into cloned proteins [2,3]. Moreover, significant advances in synthetic chemistry have recently transformed our ability to make proteins of small to moderate size in a practical fashion by total synthesis [4–10]. Total chemical synthesis allows systematic variation of all aspects of the covalent structure of a protein, including the polypeptide backbone itself and the stereochemistry of the chiral carbon centers. These features make chemical synthesis an important complement to site-directed mutagenesis in the study of smaller proteins and protein domains.

In many cases, the particular amino-acid residues involved in a protein's function are simply not known. One solution to this problem has been to use more exhaustive mutagenesis strategies involving the systematic replacement of each amino acid in a sequence with another residue, such as alanine [11–14]. These scanning mutagenesis approaches provide a functional profile across the region being modified. However, even scanning mutagenesis is constrained in a practical sense by the large amount of effort required to construct and analyze individual mutant proteins, regardless of whether chemical synthesis or recombinant expression is being used.

The recent emergence of combinatorial approaches in protein molecular biology has provided an alternative route to exhaustive mutagenesis of a protein sequence. By combining random oligonucleotide synthesis with protein expression in bacteria [15] or on phage [16], it is possible to rapidly generate large libraries of protein sequences. When used with appropriate screening procedures these combinatorial techniques have proven useful in addressing protein-folding questions [17,18]. Combinatorial approaches of this type are, however, also constrained to the 20 naturally occuring amino acids, and it is still necessary to individually identify selected proteins following the screening assay. Furthermore, as only a tiny fraction of the total number of proteins generated can be fully characterized, it is very difficult to obtain functional profiles over a protein sequence using these techniques.

We have recently developed a new technique for exploring the chemical basis of peptide and protein function (P.E.D., M.C.F., T.W.M. & S.B.H.K., submitted). The principle of 'protein signature analysis' is illustrated in Figure 1. First, total chemical synthesis is used to prepare an array of self-encoded proteins in which an analog unit is systematically placed throughout a region of interest in the polypeptide chain, such that each member of the array contains a single copy of the analog unit at a unique and defined position. Second, the array of synthetic proteins is subjected to selection based on a specific functional property; this results in division of the original mixture into a positive (functional) pool and a negative (non-functional) pool. In the third and final step, the identities of the protein-analog molecules and the position of the analog unit within each protein are simultaneously determined using a chemical readout system expressly built into the molecule for that purpose. The resulting decoded patterns form a signature relating the changes in chemical structure of the molecule to effects on protein function.

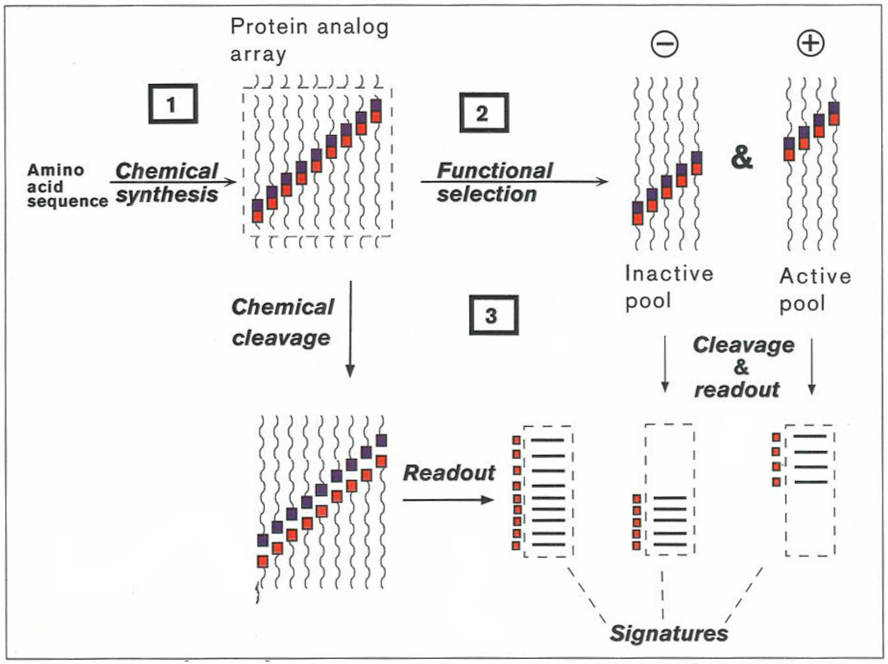
The protein signature analysis technique combines the almost unlimited versatility of chemical synthesis for systematic modification of a protein's covalent structure with the practical convenience that accompanies combinatorial methods. An important feature of the protein signature analysis approach is that it permits the synthesis, selection and readout of defined mixtures of protein analogs. This property should be contrasted with other combinatorial techniques used to probe protein structure–function relationships, for example phage display, where readout of selected clones is carried out on an individual basis. The protein signature analysis principle is completely new to protein science, although a conceptually similar strategy has been devised for studying nucleic acids [19–22].

Results and discussion

Protein signature analysis of an SH3 domain

In the form shown in Figure 1, protein signature analysis is a particularly useful way of looking at the chemical basis of ligand-binding activity in proteins. As an example of this, we have applied protein signature analysis to one of the Src homology 3 (SH3) binding domains commonly found in proteins involved in intracellular signal transduction.

Figure 1



The principle of protein signature analysis. Step 1: total chemical synthesis is used to generate an array of protein molecules derived from a single amino-acid sequence. An analog chemical structure (represented by the red and blue rectangles) is systematically incorporated at defined positions in the polypeptide chain. Step 2: the array of protein analogs is subjected to functional selection, resulting in separation into two populations: active and inactive. Step 3: the composition of each pool of analogs is then determined in a single step using a chemical readout system expressly built into the molecule for that purpose. This provides a signature relating the effects on function to substitution of the analog structure throughout the region of interest in the protein molecule.

SH3 domains are small protein modules: polypeptide chains of about 60 amino acid residues that fold to form a unique three-dimensional structure, even outside the context of the longer polypeptide chain of the parent protein. It is now well established that SH3 domains mediate protein–protein interactions through the recognition of short proline-rich sequences [23]. The high-resolution structures of several SH3 domains have been determined by NMR spectroscopy and X-ray crystallography, both in the presence and absence of bound ligand [24]. Such studies have described the folded structure of the SH3 domain, and defined many of the structural criteria governing the specific recognition of proline-rich ligands by the protein domain [25].

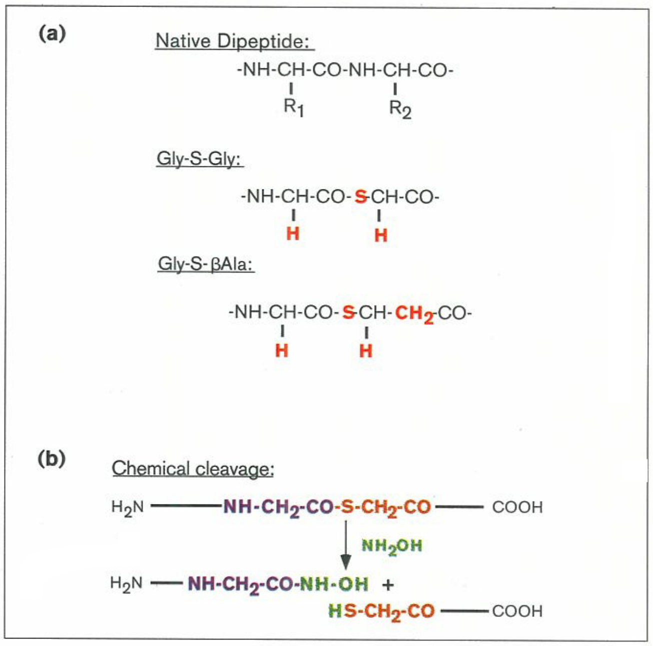
Our goal was to investigate the chemical basis of the interaction between the amino-terminal SH3 domain from the cellular adaptor protein c-Crk (residues 134–191 of the murine sequence) and its target ligand, a proline-rich peptide from the guanine nucleotide exchange protein C3G [26]. We wanted to change the chemical structure of the SH3 polypeptide chain and observe how this affected the functional properties of the domain.

Folded globular proteins found in nature are very robust, often retaining the native fold of the polypeptide chain as closely as possible, even in the face of extensive changes in the amino-acid sequence [12,27]. For this reason, we decided to introduce a dramatic perturbation in the chemical structure of the protein molecule: deletion of the sidechains of two adjacent amino acids in concert with the introduction of an extra backbone methylene. A thioester bond was also introduced to facilitate the identification of protein analogs (see below). The resulting Gly-[COS]-βAla dipeptide analog unit can be compared with a native dipeptide sequence (see Fig. 2a).

Synthesis of an array of SH3 analogs

We initially focused on a sequence of 20 amino acids near the middle of the SH3 polypeptide chain, c-Crk(146–165). As a first step an array of 19 protein analogs was chemically synthesized by placing the Gly-[COS]-βAla dipeptide unit at each possible dipeptide position within the 20-amino acid stretch. In this synthesis, we used the modified stepwise solid-phase peptide synthesis (SPPS) approach described in detail elsewhere (P.E.D., M.C.F., T.W.M. & S.B.H.K., submitted). This method made it possible to prepare all members of this array of analogs simultaneously in the course of a single synthesis (see Fig. 3). This modified split-resin procedure ensured that each individual polypeptide chain in the final product mixture contained only one analog unit at a single defined position. Stepwise synthesis of the full-length 58-residue SH3 domain polypeptide, with the introduction of a chemical perturbation at 19 defined positions of the polypeptide chain according to such a split-resin process, gave an array of analogs as a single product mixture containing the 19 desired molecular species.

Figure 2



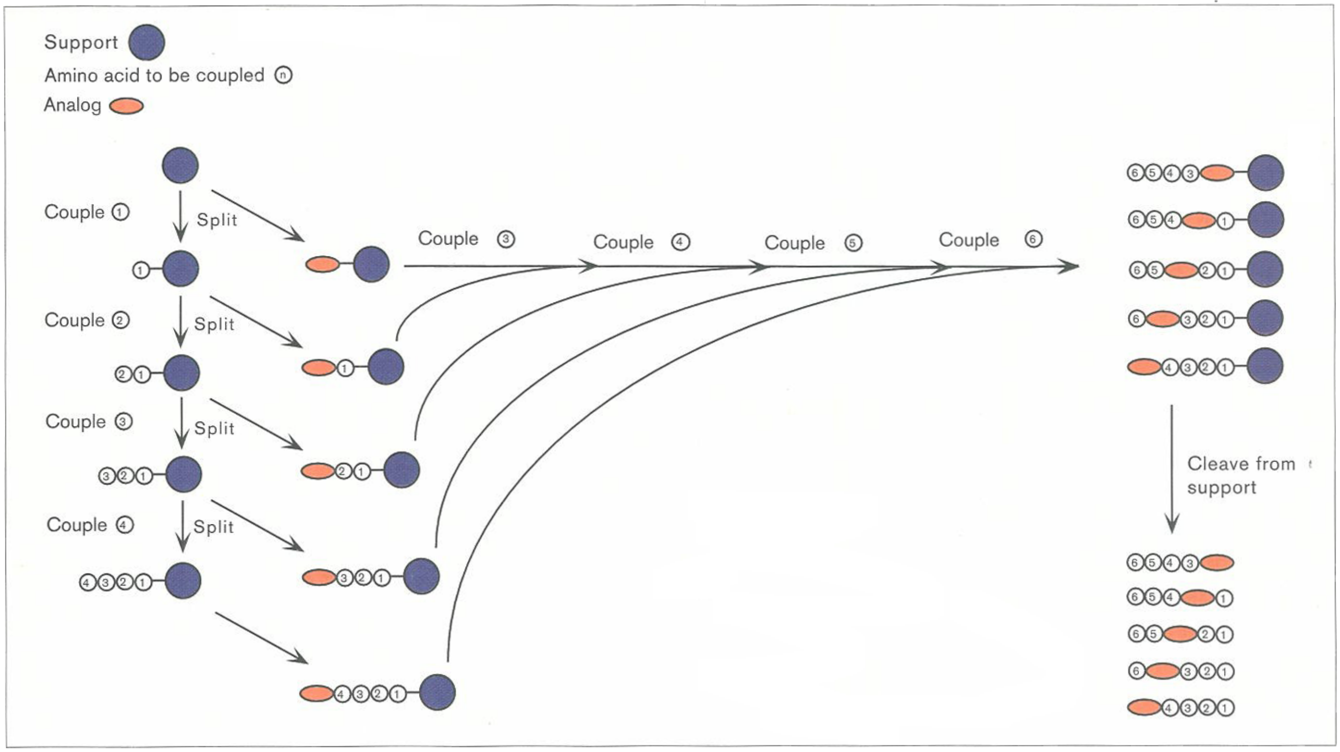
Chemical structures of analog units. **(a)** Comparison of a native dipeptide unit (top) with the structures of the Gly-[COS]-Gly (middle) and Gly-[COS]-βAla (bottom) analog units used in this study. **(b)** Selective chemical cleavage of the thioester bond within the analog unit can be carried out under mild conditions by treatment with hydroxylamine at neutral pH. The thioester bond is stable to the conditions normally used to study proteins.

Functional selection by affinity chromatography

The synthetic products were next subjected to functional selection. The amino-terminal SH3 domain from c-Crk specifically recognizes a 10-residue proline-rich sequence from the protein C3G [26]. A synthetic peptide containing the proline-rich C3G sequence was covalently immobilized on commercially available derivatized agarose beads. In preliminary experiments, a synthetic SH3 domain corresponding to the wild-type sequence was found to bind specifically to the C3G peptide affinity column, but not to a control affinity column containing an immobilized non-specific peptide sequence, corresponding to leucine enkephalin. These procedures are described in detail elsewhere (P.E.D., T.W.M., M.C.F. & S.B.H.K., submitted), and served to establish the validity of the affinity column assay.

The effects of the dipeptide analog units on the binding properties of the SH3 domain were evaluated by passing the 19-member array of synthetic analogs of the 58-residue domain, as one pool, over the C3G peptide affinity column. The non-specifically bound protein analogs (the

Figure 3



Chemical synthesis of an array of protein analogs. A modified solid-phase peptide synthesis approach was developed that made it possible to prepare all members of the array of protein analogs concurrently in the course of a single synthesis (P.E.D., M.C.F., T.W.M., & S.B.H.K., submitted). This simple split-resin procedure involves the use of two reaction vessels. At each stage of the synthesis a small aliquot of the resin (with attached peptide) is removed from the first vessel, and the analog moiety attached to the growing peptide chain. The resin aliquot is then transferred to the second reaction vessel and the remainder of the amino acids in the sequence are coupled. Continual siphoning of resin aliquots from vessel #1 into vessel #2 (with analog attachment in between), results in the generation of the complete protein array as a single product mixture. Use of this split-resin procedure ensures that each component of the array contains only a single copy of the analog at a unique and defined position.

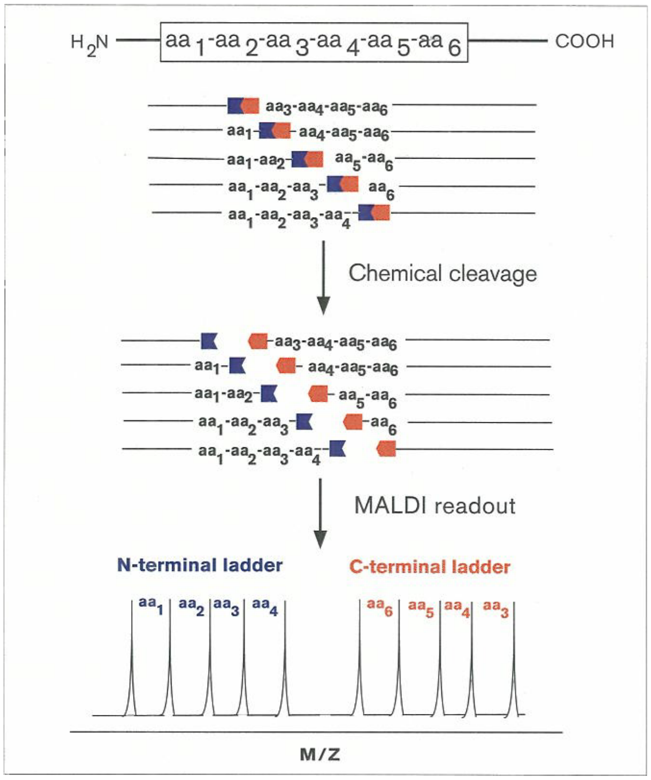
inactive pool) were first eluted from the column by washing with high-salt buffers. The remaining specifically bound protein analogs (the active pool) were then eluted by washing the affinity column with a buffer system containing 1 M NH_2OH. This buffer disrupts the specific protein–ligand interactions by chemically cleaving the protein analogs (see later) resulting in their fragmentation and hence elution from the column.

Readout of self-encoded arrays of protein analogs

The final step was to determine which protein analogs were present in each of the two pools. The identification of individual molecular species in a pool of closely related protein analogs is a formidable analytical challenge. One way to determine the molecular composition of such a mixture of protein analogs is to combine mass spectrometry with the synthetic chemistry approach, as schematically illustrated in Figure 4.

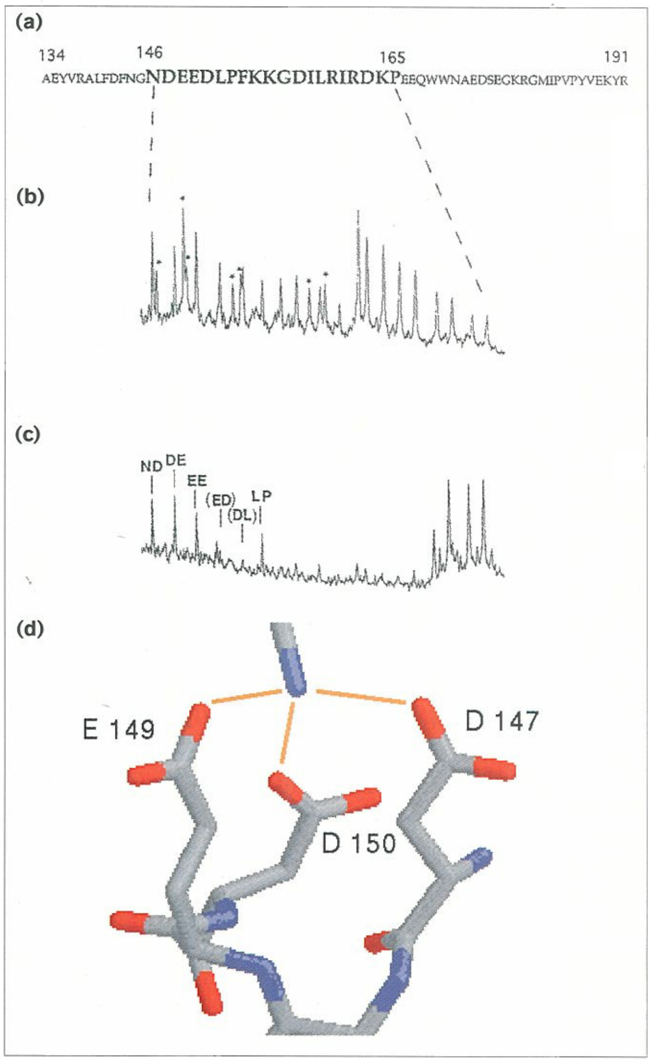
The readout of all members of each pool of protein analogs was accomplished in a single step using a chemical decoding approach, similar in concept to that already described for use with nucleic acid libraries [19–22]. A latent readout chemistry was built into each molecule in the course of the preparation of the protein array by total chemical synthesis. The analog unit contained a unique thioester chemical cleavage site (see Fig. 2b), which allowed us to chemoselectively 'unzip' the mixture of analog-containing polypeptide chains found in either the binding or non-binding pool. When examined by matrix-assisted laser desorption ionization (MALDI) mass spectrometry [28], the resulting sets of 'decoded' peptide fragments gave characteristic signatures that could be interpreted as follows. Each component of the mass spectrometric signature reflected the presence of the corresponding full-length polypeptide chain (containing the analog unit) in that pool of intact protein analogs (P.E.D., M.C.F., T.W.M. & S.B.H.K., submitted). Furthermore, the position of the analog unit in the original 58-residue c-Crk SH3 protein analog was defined by the position of the corresponding signal in the mass spectrometric signature, as schematically illustrated in Figure 4 [29,30].

Figure 4



Readout of self-encoded protein analog arrays. Building a latent chemical cleavage site into the analog unit (rectangle) means that each protein in the array will contain this chemical marker at a unique position in the polypeptide sequence. Chemical cleavage specifically at the analog unit gives rise to characteristic peptide fragments, each with a unique mass indicative of the position of the analog unit within the sequence of the original protein analog. Each protein analog in an array is thus self-encoded. Readout of these decoded peptide fragments can then be performed, in one operation, using MALDI mass spectrometry [29].

Figure 5



Protein signature analysis of a 20-residue region of the amino-terminal SH3 domain of murine c-Crk. **(a)** The highlighted amino acid sequence (residues 146–165) was substituted by the dipeptide analog Gly-[COS]-βAla, giving a 19-member array of synthetic analog proteins. **(b)** Signature of parent array after cleavage with neutral NH_2OH and analysis by MALDI mass spectrometry. Only the family of amino-terminal fragments was observed in the spectrum. The carboxy-terminal fragments, although necessarily present, are not visible under the conditions used. Asterisks indicate terminated peptides arising from impurities in the commercial amino acids used. **(c)** Signature of the active (binding) pool eluted from the C3G-derived synthetic peptide affinity column. Eight analogs showed binding under these conditions; thus, the dipeptide sequences N146D147, D147E148, E148E149, L151P152, I161R162, R162D163, D163K164, and K164P165 could be replaced by the dipeptide analog without significant loss of activity. Dipeptide sequences in parentheses ((ED) and (DL)), indicate the Gly-[COS]-βAla-containing protein analogs that do not show significant binding activity, as discussed in the text. **(d)** A region of the three-dimensional structure of the c-Crk–C3G complex [32], showing the three acidic residues within the RT loop of the SH3 domain interacting with Lys8 of the bound C3G peptide ligand. These interactions are believed to make an important contribution to binding, and to be critical in orienting the interaction of c-Crk with C3G.

Role of SH3 sidechains.

The three components of protein signature analysis — synthesis, selection, and readout — will be described in detail in a series of model studies (P.E.D., M.C.F., T.W.M. & S.B.H.K., submitted). Here we have applied the integrated approach to an SH3 domain to elucidate the chemical basis of ligand binding. The results obtained from applying functional selection/chemical readout to the 19-member array of protein analogs corresponding to the amino-terminal SH3 domain from c-Crk are shown in Figure 5. The mixture of synthetic protein analogs was passed over a C3G-peptide affinity column to assay for binding activity. The signature of the parent array of protein analogs (Fig. 5b) is compared with the signature of the pool that showed binding activity (Fig. 5c).

Eight (out of 19) members of the array of protein analogs bound to the C3G peptide. Perhaps of most interest was the pattern of binding and non-binding observed for proteins modified within the c-Crk(146–152) region. This sequence of the SH3 protein corresponds to the so-called 'RT loop', a region known to be involved in ligand binding throughout the SH3 domain family [24]. The sequence of this part of the c-Crk SH3 polypeptide is Asn-Asp-Glu-Glu-Asp-Leu-Pro. It is evident from the signature of the functional pool of SH3 analogs (Fig. 5c) that binding to the C3G-derived peptide ligand occurred even when the sidechains of the Asp147 or Glu149 residues in the SH3 domain had been removed. Thus, interaction with the Asp147 or Glu149 sidechain carboxyls was not essential for binding (note that the effect of the replacement of these two residues on binding specificity cannot be inferred from this experiment). In contrast, removal of the sidechain of Asp150 by substitution with either the Gly or βAla portion of the dipeptide analog unit virtually eliminated binding; restoration of Asp150 restored binding. It should be noted that Asp147 and Asp150 are both conserved in the viral form of the protein, v-Crk, whereas Glu149 is replaced with Gly [31].

These data are intriguing and offer experimental support for a difference in the roles of the three acidic sidechains in ligand binding, as previously suggested by the X-ray crystallographic data [32]. As shown in Figure 5d, all three of the sidechain carboxylate functionalities in residues Asp147, Glu149, and Asp150 of the c-Crk SH3 domain make specific interactions with the sidechain -ϵNH_3^+ of Lys8 in the ligand. From the protein signature analysis results presented here we can infer that the primary determinant of binding in this region of the SH3 molecule is the Asp150 sidechain carboxylate.

The interaction of Asp147 and Glu149 sidechains with the peptide ligand may have a different role, perhaps affecting the specificity of binding by discriminating between Lys and Arg sidechains at this position [32]. The predominant role of Asp150 and the different roles of Asp147 and Glu149 were both suggested by the crystallography data. In the crystal structure, the -ϵNH_3^+ group of the lysine residue (in the Pro-rich peptide ligand) forms a hydrogen bond to an oxygen atom in the sidechain carboxylate of Asp150, using the preferred *syn* orientation of the oxygen lone electron pair as shown (Fig. 5d), whereas the hydrogen bonds to Asp147 and Glu149 are in the less favored *anti* orientation. The signature analysis data shown in Figure 5c are consistent with this crystallographic observation, because replacement of Asp150 resulted in gross loss of binding activity, whereas replacement of Asp147 or Glu149 did not.

Productive application of the signature analysis approach does not require knowledge of the three-dimensional structure of a protein domain. However, the three-dimensional structure can be used together with protein signature analysis to give additional insights into the chemical basis of protein function. In the example given here, the combination of signature analysis data with structural studies gave a more informative interpretation of the molecular basis of ligand binding than would have been possible with protein signature data alone. These results also show the potential of the protein signature analysis technique to illuminate the chemical reality of mechanisms suggested by the structural data.

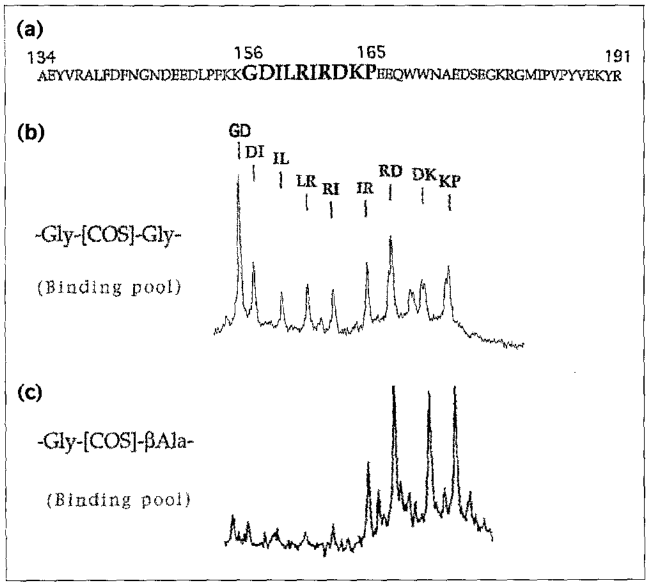
Role of SH3 backbone

As shown in Figure 5, eight of the 19 protein analogs displayed binding activity, while 11 were inactive. We have discussed the implications of these observations for the roles of specific amino acid sidechains (above). How can we use these data to obtain information on the other factors that determine the chemical basis of binding by this SH3 domain? The non-functional members of the array of protein analogs could be inactive for any or all of the following reasons: deletion of amino acid sidechains, insertion of an extra methylene in the polypeptide backbone, and deletion of the hydrogen-bonding ability of the central amide moiety in the analog structure. Information on which of these mechanisms is responsible for the loss of activity can be simply obtained by making another array containing alternative analog structures covering the region of interest in the SH3 domain. In this case, we made a second nine-membered array over the region cCrk(156–165) using -Gly-[COS]-Gly- as an analog unit in which the additional methylene of the original analog unit was not present (Fig. 2a).

The signature obtained after functional separation, based on binding to the proline-rich C3G peptide affinity column, and readout of this new nine-membered array of protein analogs is shown in Figure 6b. The signature in Figure 6c represents an expansion of the signature data shown in Figure 5c, but focussing on the region corresponding to replacement of residues 156–165 of the SH3 domain by the Gly-[COS]-βAla analog. Of the nine Gly-[COS]-βAla-containing SH3 analogs in this region, only four bound to the C3G peptide affinity column. The other five analogs did not bind under the conditions used. This pattern is identical to that observed for this region in a data set in which only these nine analogs were analyzed (P.E.D., T.W.M., M.C.F. & S.B.H.K., submitted).

In contrast, all nine (-Gly-[COS]-Gly-)-containing protein analogs exhibited appreciable binding activity in an identical assay (Fig. 6b). That so many of the protein analogs retained specific binding activity is remarkable, given the very substantial nature of the chemical changes made in the polypeptide chain. The data show that neither the pairwise deletion of sidechains nor deletion of the hydrogen bond in the central amide moiety of the analog structure was responsible for the lack of binding exhibited by the five inactive members of the original -Gly-[COS]-

Figure 6



Iterative protein signature analysis applied to the amino-terminal SH3 domain of murine c-Crk. **(a)** The amino acid sequence of the 58-residue polypeptide chain. Protein signature analysis was used to study how chemical variation of the central 10-residue region (highlighted residues 156–165) affected C3G peptide binding. Two rounds of signature analysis were performed, using different dipeptide analog units. **(b)** Round 1. Signature of the active (binding) pool obtained from the nine-membered array of Gly-[COS]-Gly-containing protein analogs. In contrast to the previous experiment, analysis of this signature reveals that all nine protein analogs were present in the binding pool. **(c)** Round 2. Signature of the active (binding) pool resulting from passing the parent array of Gly-[COS]-βAla-containing protein analogs over a C3G-derived synthetic peptide affinity column. The signature data shown represents an expansion of the larger signature shown in Figure 5c. Only four dipeptide sequences out of a total of nine (I161R162; R162D163; D163K164; and K164P165) in this region could be replaced by Gly-[COS]-βAla without significant loss of binding activity.

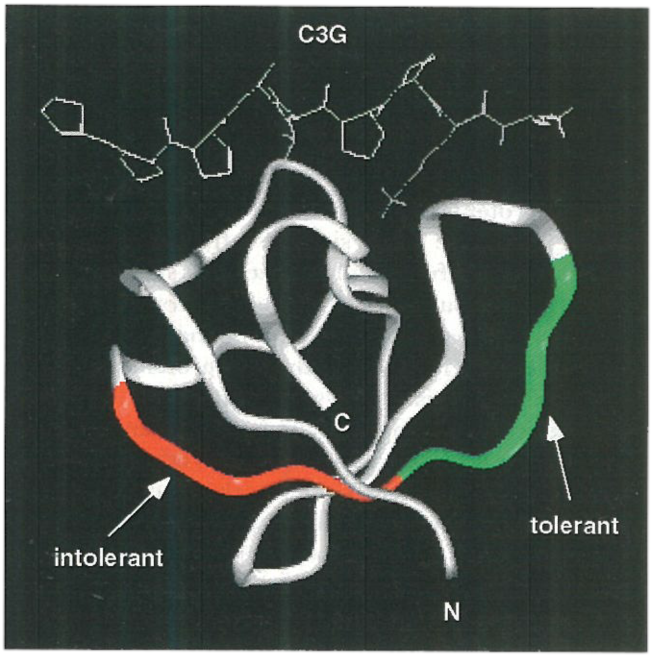
βAla-containing array of protein analogs covering this region. Rather, it can be inferred from comparison of the two sets of data that the observed lack of binding activity was caused by insertion of the extra methylene in the polypeptide backbone by the original analog unit. Thus, it appears that the region defined by residues c-Crk(156–161) is less tolerant to backbone engineering than the region defined by residues c-Crk(161–165) (Fig. 7). The affinity binding assay used does not discriminate between a gross structural pertubation and a purely functional effect, since both could result in a loss of activity. However, none of the amino acids in the region being studied (residues 156–165) interact directly with the ligand, suggesting that the observed effects may be structural in origin.

Implications

These results clearly demonstrate the power of the protein signature analysis approach. These data could only have been obtained using a chemistry-driven technique of this type. Thus, precise dissection of the molecular basis of effects on protein function is possible through iterative application of the signature analysis method. Information obtained at each stage of studying a system can be used to guide systematic elucidation of the chemical basis of function for an entire molecule. At each cycle of experimentation, the signatures are correlated with chemical changes at all sites being studied in the protein molecule. Such a process will inevitably reveal the chemical basis of the effects on activity.

What contribution can the techniques described in this paper make to understanding how proteins work? Advances in total synthesis have opened the world of proteins, and in particular protein domains, to the tools of organic chemistry [33,34], allowing for variations of molecular structure not constrained by the genetic code (for example [35,36]). Total chemical synthesis thus provides an important complement to, and extension of, the powerful recombinant DNA-based techniques commonly used to study the molecular basis of protein function by variation of amino acid sidechains [1,11,15,16]. Using chemical synthesis it is possible to alter the covalent structure of a protein in ways simply not accessible to current molecular biology techniques, despite improvements of the recombinant DNA approach which have

Figure 7



Results from two rounds of protein signature analysis of the sequence comprising residues 156–165 superimposed on the crystal structure of the amino-terminal c-Crk SH3 domain complexed to the proline-rich C3G peptide [32]. Indicated are those regions of the polypeptide chain observed to be either tolerant (green) or intolerant (red) of an extra backbone methylene group.

allowed for a variety of non-coded elements to be introduced into individual proteins at single sites [2,3]. However, the systematic application of both recombinant and chemical techniques has been tempered by the effort required to make, characterize, and assay individual analog proteins. What is needed is an approach that combines the versatility of chemistry with the ability to obtain, in a practical fashion, a qualitative profile of the functional effects of structure variation throughout the protein molecule.

Protein signature analysis is just such an approach. In essence it provides an empirical signature relating defined changes in the chemical structure of a protein to effects on its biological function. Synthesis of arrays of analog proteins allows the straightforward exploration of the entire sequence of a functional domain, or other region of interest in a protein molecule. Furthermore, modular chemical ligation [37,38] will allow the study of functional domains in the context of larger molecules, giving a global picture of the chemical basis of protein function.

Significance

Altering the covalent structure of a protein is a powerful way of elucidating the chemical basis of function. Here we describe a new technique, protein signature analysis, that allows the qualitative effect of performing a defined chemical modification to be evaluated simultaneously at all positions within a sequence. Protein signature analysis combines the versatility of chemistry for the systematic modification of a protein's covalent structure with the practical convenience that accompanies combinatorial synthesis. This combinatorial protein chemistry approach has been applied to the amino-terminal SH3 domain from c-Crk. A total of 28 chemically defined protein analogs were analyzed in only two protein signature analysis experiments. Using protein signature analysis, the effect on biological function of modifying both amino acid sidechains and the polypeptide backbone of the protein was determined. The latter of these, systematic backbone engineering, is unprecedented in the study of proteins. Protein signature analysis allows fundamental biological processes such as protein folding, binding and catalysis to come under the scrutiny of synthetic organic chemistry. Consequently, the technique provides a framework for the systematic application of chemistry to deciphering how proteins work, and thus complements the analogous chemical approaches already available for the study of nucleic acids [22].

Materials and methods

Synthesis of peptides

With the exception of protein arrays (see below) all peptides were chemically synthesized according to optimized solid-phase methods [4], and purified by preparative reverse-phase high-pressure liquid chromatography (HPLC) using a Vydac C-18 column. In all cases, peptide composition and purity were confirmed by electrospray mass spectrometry and analytical reverse-phase HPLC.

Synthesis of c-Crk SH3 protein arrays

A detailed description of the split-resin procedure used will be published elsewhere (P.E.D., M.C.F., T.W.M. & S.B.H.K., submitted). Briefly, the technique involves the use of two reaction vessels with identical synthetic manipulations being carried out in each. Standard stepwise chain assembly [4] was initiated on resin in the first vessel (0.2 mmole scale); peptide–resin samples were repeatedly removed from the first vessel at each stage of the synthesis, and analog units were introduced into the polypeptide chain by coupling as preformed HOBt esters; after modification, the samples were transferred to the second vessel for completion of the chain assembly by standard stepwise chain assembly. The size of the samples was adjusted to yield approximately equal molar amounts of each protein analog in the array (dependent on the number of protein analogs in a given array). The dipeptide analogs Gly-[COS]-βAla and Gly-[COS]-Gly were prepared as previously described [39]. Upon completion of the synthesis, each parent protein array was characterized as follows: crude protein array (~1 mg) was dissolved in a cleavage buffer consisting of 1 M NH_2OH, 20 mM NH_4HCO_3, pH 6.5 buffer (1 ml) and stirred for 15 min. The cleaved arrays were then exchanged into a 70 % CH_3CN:30 % H_2O, 0.1 % trifluoroacetic acid (TFA) solvent system (using a 1 ml C-18 desalting column) and immediately analyzed by MALDI mass spectrometry.

Synthesis of peptide affinity columns

The C3G peptide affinity column was prepared as follows. The peptide Ac-CWBPPPALPPKKR-NH_2 (B = ε-aminocaproic acid) was dissolved in 50 mM Tris, 5 mM EDTA, pH 8.0 (10 mg in 2 ml) and shaken with Sulfolink™ resin (Pierce) for 1 h. Unreacted iodoalkyl groups on the resin were then blocked by treatment with 50 mM cysteamine, 50 mM Tris, 5 mM EDTA, pH 8.0 buffer. The loading of the column was determined by UV to be ~5 μmole ml^{-1}. A similar procedure was used to attach the control peptide Ac-CBYGGFL-NH_2 (YGGFL = leucine enkephalin) to the Sulfolink™ support.

Affinity selection of protein arrays

The crude protein array (1.5 mg) was dissolved in 20 mM Hepes, 50 mM NaCl pH 7.3 buffer (600 μl) and loaded onto a 1-ml C3G peptide affinity column pre-equilibrated with the same buffer. After 6–8 h the non-specifically bound material was eluted from the column by washing with 0.5 M NaCl, 0.1 M sodium phosphate, pH 7.0 buffer (6 x 1 ml). Eluted material (typically in the first and second wash) was immediately cleaved by dilution into 1 M NH_2OH, 20 mM NH_4HCO_3, pH 6.5 buffer. Following further column washing with 1 M NaCl, 0.1 M sodium phosphate, pH 7.0 buffer, the specifically bound material (active pool) was chemically cleaved and simultaneously eluted from the affinity column by washing with 1 M NH_2OH, 20 mM NH_4HCO_3, pH 6.5 buffer (4 x 1 ml). Eluted fractions were exchanged into a 70 % CH_3CN:30 % H_2O, 0.1 % TFA solvent system and immediately analyzed by MALDI mass spectrometry.

MALDI analysis of peptide arrays

All samples were prepared by adding 2 μl of the desalted column fraction to 5 μl of a saturated solution of α-cyano cinnaminic acid in 50 % acetonitrile in water, 0.1 % TFA. From this mixture, 2 μl, containing ~1–10 pmole of each pepide component, was added to a stainless steel probe tip and the solvent allowed to evaporate under ambient conditions. Mass spectra were recorded using a prototype laser desorption, linear time-of-flight mass spectrometer from Ciphergen Biosystems (Palo Alto, CA). Samples were desorbed/ionized using 337 nm radiation output from a nitrogen laser (Laser Science, Inc., Newton MA). All spectra were acquired in the positive-ion mode and summed over 20–50 laser pulses. Time-to-mass conversion was accomplished by internal calibration using the $[M+H]^+$ signals from the largest and smallest peptide components in each array.

Acknowledgements

We thank Michael Youngquist and Steve Clark for helpful discussions regarding mass spectrometry data reduction. We thank Ciphergen Biosystems (Palo Alto, CA) for their generous gift of a MALDI mass spectrometer. Studies reported here were funded in part by NIH grants GM48870, HL31950, and GM48897. M.C.F. was an Amgen Fellow.

References

1. Smith, M. (1994), Synthetic DNA and biology. (Nobel lecture). *Angew. Chem. Int. Ed. Engl.* **33**, 1214–1221.
2. Mendel, D., Ellman, J.A., Chang, Z., Veenstra, D.L., Kollman P.A. & Schultz, P.G. (1992). Probing protein stability with unnatural amino acids. *Science* **256**, 1798–1802.
3. Cornish, V.W., Mendel, D. & Schultz, P.G. (1995). Probing protein structure and function with an expanded genetic code. *Angew. Chem. Int. Ed. Engl.* **34**, 621–633.
4. Schnölzer, M., Alewood, P., Jones, A., Alewood D. & Kent, S.B.H. (1992). *In situ* neutralization in Boc-chemistry solid phase peptide synthesis: rapid, high yield assembly of difficult sequences. *Int. J. Pept. Protein Res.* **40**, 180–193.
5. Schnölzer, M. & Kent, S.B.H. (1992). Constructing proteins by dovetailing unprotected synthetic peptides: backbone-engineered HIV protease. *Science* **256**, 221–225.
6. Gaertner, H.F., Rose, K., Cotton, R., Timms, D., Camble R. & Offord, R.E. (1992). Construction of protein analogues by site-specific condensation of unprotected peptides. *Bioconjugate Chem.* **3**, 262–268.
7. Dawson, P.E., Muir, T.W., Clark-Lewis I. & Kent, S.B.H. (1994). Synthesis of proteins by native chemical ligation. *Science* **266**, 776–779.
8. Kemp, D.S. & Carey, R.I. (1993). Synthesis of a 39-peptide and a 25-peptide by thiol-capture ligations: observation of a 40-fold rate acceleration of the intramolecular O,N-acyl transfer reaction between peptide fragments bearing only cysteine protecting groups. *J. Org. Chem.* **58**, 2216–2222.
9. Liu, C.-F. & Tam, J.P. (1994). Peptide segment ligation strategy without use of protecting groups. *Proc. Natl. Acad. Sci. USA* **91**, 6584–6588.
10. Nefzi, A., Sun X. & Mutter, M. (1995). Chemoselective ligation of multifunctional peptides to topological templates via thioether formation for TASP synthesis. *Tetrahedron Lett.* **36**, 229–230.
11. Cunningham, B.C. & Wells, J.A. (1989). High-resolution epitope mapping of hGH–receptor interactions by alanine-scanning mutagenesis. *Science* **244**, 1081–1085.
12. Blaber, M., Baase, W.A., Gassner N. & Matthews, B.W. (1995). Alanine-scanning mutagenesis of the α-helix 115–123 of phage T4 lysoyme: effect on structure, stability and binding of solvent. *J. Mol. Biol.* **246**, 317–330.
13. Tam, J.P., *et al.*, & Ke, X.-H. (1989). Systematic approach to study the structure–activity of transforming growth factor α. In *Peptides: chemistry and biology* [Proceedings of the 11th American Peptide Symposium.] pp. 75–77, ESCOM, Leiden.
14. Gibbs, C.S., *et al.*, & Leung, L.L.K. (1995). Conversion of thrombin into an anticoagulant by protein engineering. *Nature* **378**, 413–416.
15. Reidhaar-Olsen, J.F. & Sauer, R.T. (1988). Combinatorial cassette mutagenesis as a probe of the informational content of protein sequences. *Science* **241**, 53–57.
16. Scott, J.K. & Smith, G.P. (1990). Searching for peptide ligands using an epitope ligand. *Science* **249**, 386–390.
17. Davidson, A.R., Lumb K.L. & Sauer, R.T. (1995). Cooperative folding proteins in random sequence libraries. *Nat. Struct. Biol.* **2**, 856–864.
18. Kamteker, S., Schiffer, J.M., Xiong, H., Babik, J.M. & Hecht, M.H. (1993). Protein design by binary patterning of polar and non-polar amino acids. *Science* **262**, 1680–1685.
19. Hayashibara, K.C. & Verdine, G.L. (1991). Template-directed interference footprinting of protein–guanine contacts in DNA. *J. Am. Chem. Soc.* **113**, 5104–5106.
20. Hayashibara, K.C. & Verdine, G.L. (1992). Template-directed interference footprinting of cytosine contacts in a protein–DNA complex: potent interference of 5-aza-2'-deoxycytidine. *Biochemistry* **31**, 11265–11273.
21. Mascarenas, J.L., Hayashibara, K.C. & Verdine, G.L. (1993). Template-directed interference footprinting of protein–thymidine contact. *J. Am. Chem. Soc.* **115**, 373–374.
22. Min, C., Cushing, T.D. & Verdine, G.L. (1996). Template-directed interference footprinting of protein–adenine contacts. *J. Am. Chem. Soc.* **118**, 6116–6120.
23. Pawson, T. & Schlessinger, J. (1993). SH2 and SH3 domains. *Curr. Biol.* **3**, 434–442.
24. Kuriyan, J. & Cowburn, D. (1993). Structures of SH2 and SH3 domains. *Curr. Opin. Struct. Biol.* **3**, 828–837.
25. Feng, S., Chen, J.K., Yu, H., Simon J.A. & Schreiber, S.A. (1994). Two binding orientations for peptides to the Src SH3 domain: development of a general model for SH3–ligand interactions. *Science* **266**, 1241–1247.
26. Knudsen, B.S., Feller S.M. & Hanafusa, H. (1994). Four proline-rich sequences of the guanine-nucleotide exchange factor, C3G, bind with unique specificity to the first Src homology 3 domain of Crk. *J. Biol. Chem.* **269**, 32781–32787.
27. Shortle, D. & Sondek, J. (1995). The emerging role of insertions and deletions in protein engineering. *Curr. Opin. Biotechnol.* **6**, 387–393.
28. Chait, B.T. & Kent, S.B.H. (1992). Weighing naked proteins: practical high accuracy mass measurement of peptides and proteins. *Science* **257**, 1885–1894.
29. Chait, B.T., Wang, R., Beavis R.C. & Kent, S.B.H. (1993). Protein ladder sequencing. *Science* **262**, 89–92.
30. Zhao, Y., Muir, T.W., Kent, S.B.H., Tischer, E., Scardina, J.M. & Chait, B.T. (1996). Mapping protein–protein interactions by affinity-directed mass spectrometry. *Proc. Natl. Acad. Sci. USA* **93**, 4020–4024.
31. Mayer, B.J. & Hanafusa, H. (1993). Mutagenic analysis of the v-crk oncogene: requirement for the SH2 and SH3 domains and correlation between increased cellular phosphotyrosine and transformation. *J. Virol.* **64**, 3581–3589.
32. Wu, X., *et al.*, & Kuriyan, J. (1995). Structural basis for the specific interaction of lysine-containing proline-rich peptides with the N-terminal SH3 domain of c-Crk. *Structure* **3**, 215–226.
33. Muir, T.W. & Kent, S.B.H. (1993). The chemical synthesis of proteins. *Curr. Opin. Biotechnol.* **4**, 420–427.
34. Muir, T.W. (1995). A chemical approach to the construction of multimeric protein assemblies. *Structure* **3**, 649–652.
35. Baca, M., Alewood, P.F. & Kent, S.B.H. (1993). Structural engineering of the HIV-1 protease molecule with a β-turn mimic of fixed geometry. *Prot. Sci.* **2**, 1085–1091.
36. Baca, M. & Kent, S.B.H. (1993). Catalytic contribution of flap-substrate hydrogen bonds in HIV-1 protease explored by chemical synthesis. *Proc. Natl. Acad. Sci. USA* **90**, 11638–11642.
37. Canne, L.E., Ferre-D'Amare, A.R., Burley, S.K. & Kent, S.B.H. (1995). Total chemical synthesis of a unique transcription factor-related protein: c-Myc–Max. *J. Am. Chem. Soc.* **117**, 2998–3007.
38. Baca, M., Muir, T.W., Schnölzer, M. & Kent, S.B.H. (1995). Chemical ligation of cysteine-containing peptides: synthesis of a 22-kDa tethered dimer HIV-1 protease. *J. Am. Chem. Soc.* **117**, 1881–1887.
39. Hojo, H. & Aimoto, S. (1991). Polypeptide synthesis using the S-alkyl thioester of a partially protected peptide segment. Synthesis of the DNA-binding domain of c-Myb protein (142–193)-NH_2. *Bull. Chem. Soc. Jpn.* **64**, 111–117.