Pergamon

Tetrahedron 56 (2000) 9461–9470

TETRAHEDRON

Synthesis of Multi-Domain Proteins Using Expressed Protein Ligation: Strategies for Segmental Isotopic Labeling of Internal Regions

Ulrich K. Blaschke, Graham J. Cotton and Tom W. Muir*

Laboratory of Synthetic Protein Chemistry, The Rockefeller University, 1230 York Avenue, New York, NY 10021, USA

Received 18 April 2000; accepted 26 June 2000

Abstract—Here we describe how a sequential version of the protein semi-synthesis technique, Expressed Protein Ligation (EPL), can be used to assemble multiple (i.e. 3 or more) recombinantly-derived polypeptides segments into a target protein. Sequential EPL was successfully used to assembly the 304 amino acid eukaryotic adaptor protein, Crk-II, from three recombinant polypeptide segments in good yield. Moreover, the resulting multi-component ligation product was found to possess the expected biological activity in a series of ligand binding studies. By allowing the controlled assembly of 3 or more recombinant polypeptide segments, sequential EPL opens the door to the segmental isotopic labeling of internal regions of large proteins with NMR probe-nuclei.

Introduction

Expressed Protein Ligation (EPL) is a protein semi-synthesis technique which allows synthetic peptides and recombinant proteins to be chemoselectively and regioselectively joined together via a normal peptide bond.[1–3] The technique combines the structural flexibility associated with chemical peptide synthesis with the extended size range of recombinant DNA expression, i.e. semi-synthetic proteins can be prepared from short synthetic peptide cassettes (containing any number of chemical probes) and much larger recombinant polypeptide building blocks. As with other approaches designed to allow the introduction of unnatural amino acids into proteins,[4–9] EPL offers the possibility of using synthetic chemistry to study protein structure and function in a manner entirely analogous to that routinely used to elucidate the structure-activity relationships (SAR) of small bioactive peptides.[10] This type of 'chemical protein mutagenesis' provides expanded opportunities in protein engineering and should have applications in areas ranging from structural biology and biochemistry to basic cell biology.

EPL is an extension of the well established native chemical ligation approach, originally developed for the total chemical synthesis of proteins from fully unprotected synthetic peptides.[7–12] Native chemical ligation is based on the chemoselective reaction of a polypeptide containing a C-terminal thioester (α-thioester) with a second polypeptide containing an N-terminal cysteine (α-Cys) residue. At neutral pH, such unprotected peptides chemoselectively react with one another to form a native peptide bond at the ligation site.[11] This is an extremely robust chemical process which can be performed in the presence of chaotropes, detergents, organic solvents and, importantly, all of the chemical functionalities found commonly in proteins. Thus, native chemical ligation is widely used in the synthesis of small proteins and protein domains.[7,12] Significantly, recent advances in protein engineering allow both the necessary reactive groups for native chemical ligation to be introduced into recombinant polypeptides (Fig. 1). This, of course, means that all possible combinations of synthetic and recombinant building blocks can be assembled in a semi-synthetic ligation process, commonly referred to as Expressed Protein Ligation.

As illustrated in Fig. 1, recombinant polypeptides containing α-Cys moieties or α-thioester groups can be generated directly from the appropriate fusion-protein precursors. Reactive recombinant protein α-thioesters can be prepared by exploiting the natural process, protein splicing, a post-translational event known to involve thioester intermediates.[13–16] It is possible to chemically intercept the splicing process with a suitable thiol by appending the recombinant polypeptide of interest to a mutated protein splicing domain (termed an intein,[17]), thereby generating the corresponding recombinant protein α-thioester (Fig. 1A). Two general strategies have been developed for the production of α-Cys containing recombinant polypeptides, both of which involve proteolytic removal of an N-terminal leader sequence from a precursor fusion protein

Keywords: semi-synthesis technique; expressed protein ligation; isotopic labeling.

* Corresponding author. Tel.: +1-212-327-7368; fax: +1-212-327-7358; e-mail: muirt@rockvax.rockefeller.edu

0040–4020/00/$ - see front matter
PII: S0040-4020(00)00830-9

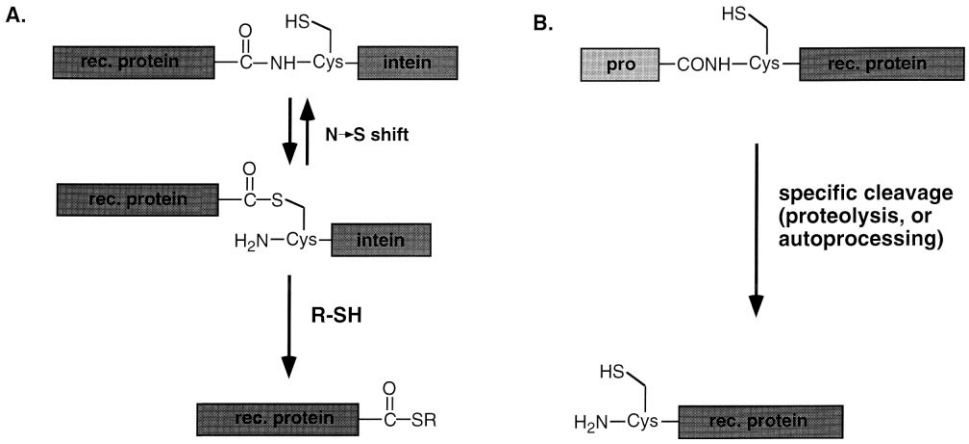

Figure 1. Generation of reactive recombinant proteins used in Expressed Protein Ligation. (A) Intein-fusion strategy used to generate recombinant protein α-thioesters. (B) Fusion protein strategy used to generate recombinant proteins possessing an N-terminal cysteine residue (α-Cys). The pro-peptide can contain the recognition sequence for a specific proteolytic enzyme, or can be an autoprocessing protein domain such as an engineered intein.

(Fig. 1B). The first approach makes use of specific proteases which cleave C-terminal to their recognition site—factor Xa, enterokinase or Ubiquitin C-terminal hydrolase. By appending the appropriate recognition sequence immediately in front of a cryptic N-terminal cysteine in the protein of interest, it is possible to generate the requisite α-Cys moiety by in vitro treatment of the fusion protein with the protease.[18] An alternative strategy has recently been described which does not require a separate proteolysis step but instead makes use of an auto-processing fusion protein system.[19–21] This approach was again developed from studies on protein splicing and utilizes yet another engineered *intein* splicing domain which spontaneously cleaves itself off the precursor fusion protein, again to give the requisite N-terminal cysteine protein.

Expressed protein ligation has been used in a variety of protein engineering studies, allowing the incorporation of post-translational modifications,[1] unnatural amino acids,[22] and biochemical/biophysical probes[2,18,23] into proteins. The technology has also been used to generate backbone cyclized and polymeric proteins,[24–27] as well as proteins which act as biosensors for signal transduction processes.[28,29] Of particular relevance to the present work, EPL and related technologies (see below) have been used in so-called 'Segmental Isotopic Labeling' strategies[30–33] designed to address the practical size limit for protein structure determination using nuclear magnetic resonance (NMR) spectroscopy. This limit is attributable to the loss of spectral resolution occurring from both increased line widths at longer rotational correlation times, and from the increased number of signals of similar chemical shifts—both effects are proportional to the number of amino acids in the protein. The former of these problems has to some extend been addressed with the development of new NMR experiments such as Transverse Relaxation Optimized Spectroscopy (TROSY)[34] and approaches for measuring residual dipolar coupling constants.[35] However, standard isotopic labeling strategies involving uniform incorporation of ^{13}C, ^{15}N and perdeuteration of amino acid side-chains do not address the signal overlap problem for larger systems. Segmental isotopic labeling seeks to resolve this issue by allowing selected segments of a protein to be isotopically labeled with ^{13}C, ^{15}N and ^{2}H. In principle, segmental isotopic labeling of a large protein will lead to simplified NMR spectra; unlabeled regions of the protein can be filtered out using suitable heteronuclear correlation experiments leaving only signals from the labeled part of the protein. Consequently, segmental labeling should allow NMR structure analysis of discrete regions of very large proteins (Fig. 2).

The feasibility of segmental isotopic labeling has recently been demonstrated using two different peptide ligation strategies, trans-splicing[30,32,33] and EPL.[31] The *trans-splicing* approach is based on the observation that protein splicing can be triggered by reconstituting inactive N- and C-terminal fragments of an intein.[36–39] This provides a means of joining any two recombinant proteins together in vitro; each protein is expressed as a fusion with complementary parts of the split-intein, simply refolding the fusion proteins together results in activation of protein splicing activity (through non-covalent association of the intein fragments) and so generation of the desired chimera. In a pioneering study, Yamazaki and co-workers[30] exploited the trans-splicing phenomenon for segmental isotopic labeling of proteins. Using a trans-splicing system based on the PI-*PfuI* intein, these researchers were able to selectively ^{15}N label the C-terminal domain of the *E. coli* RNA polymerase α subunit. Each half of the protein was selectively labeled by simply expressing the corresponding split intein fusion in ^{15}N enriched medium. Heteronuclear NMR experiments on these samples revealed clearly the improvement in spectral resolution that segmental labeling provides. In a complementary study,[31] EPL was used to ^{15}N label a single domain within a Src-homology domain pair derived from the Abl protein tyrosine kinase. An ethyl thioester derivative of recombinant Abl-SH3 was generated from the corresponding intein fusion and then chemically ligated, under physiological conditions, to ^{15}N-labeled recombinant Abl-SH2 possessing a factor Xa generated α-Cysteine. Comparison between the $^{1}H\{^{15}N\}$ NMR spectra of fully labeled and segmental labeled Abl-SH(32) again illustrated the power of segmental isotopic labeling for studying large proteins by NMR spectroscopy.

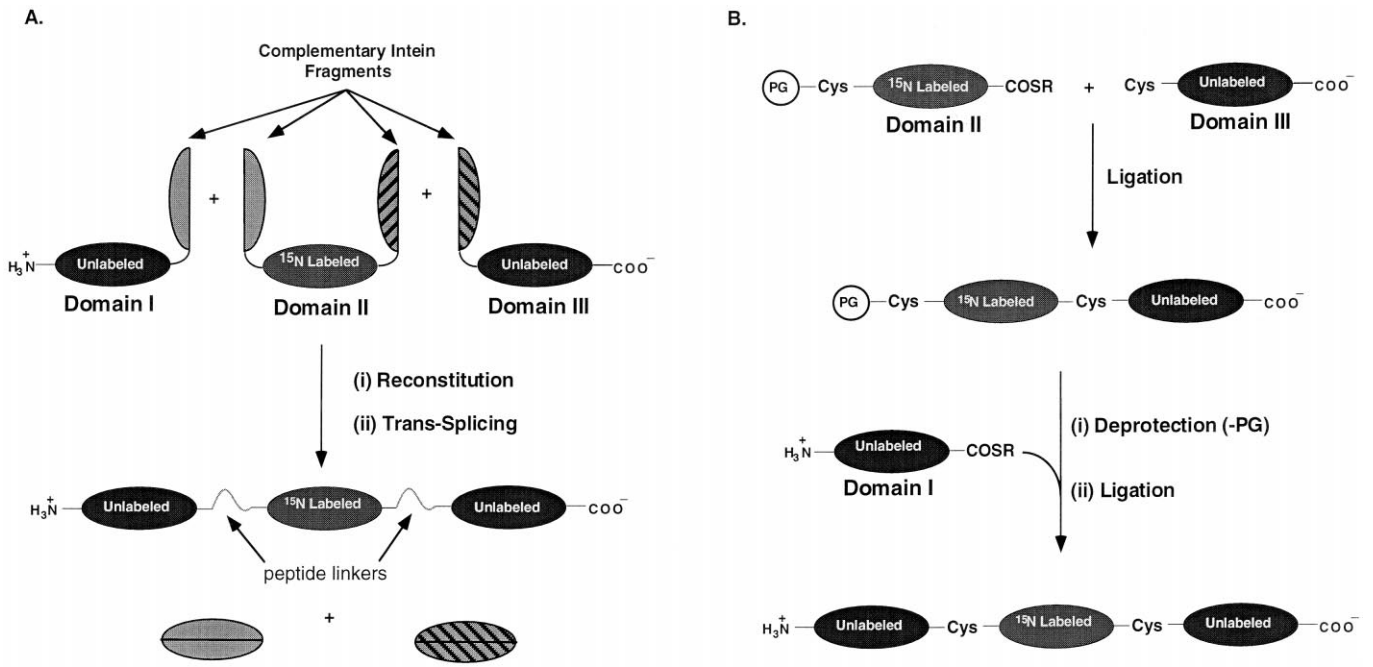

Figure 2. Strategies for isotopically labeling internal domains of a protein. (A) Split-intein approach;[33] the individual domains are expressed as split-intein fusion proteins. Using two orthogonal split-inteins (shaded and hatched hemi-ellipsoids), the assembly of the full length protein is accomplished in a one-pot reaction. This approach results in a short peptide linker being introduced at each of the two ligation sites. (B) Sequential ligation of individual domains by Expressed Protein Ligation. The individual domains are designed to carry the α-thioester and/or the cryptic N-terminal cysteine residue protected with a removable pro-sequence (PG). After a three step assembly in the C to N direction, the full length protein has only minor changes (which in the case of c-Crk-II are two Ser→Cys mutations) in the primary sequence.

The biosynthetic strategies described above allow the amino or carboxyl terminal half of a protein to be selectively isotopically labeled. A logical extension of this work would be to develop strategies which allow segmental isotopic labeling of internal regions of a large protein (Fig. 2). This important technical advance has recently been achieved by Yamazaki and co-workers[33] who selectively labeled an internal region of the model protein, Maltose Binding Protein (MBP), using a clever tandem trans-splicing method. Key to this approach was the use of two orthogonal split inteins, PI-*pfuI* and PI-*pfuII*, both from the organism *Pyrococcus furiosus*. The target MBP protein was expressed as three split-intein fusions; the central MBP segment contained PI-*pfuII* and PI-*pfuI* intein fragments at its N- and C-terminus respectively, while the N- and C-terminal MBP segments contained the complementary PI-*pfuII* and PI-*pfuI* fragments at their C- and N-termini, respectively. Synthesis of the desired internally labeled MBP was facilitated by reconstitution of the three purified polypeptides (Fig. 2A). Importantly, the authors were able to prevent undesirable inter-intein trans-splicing by splitting the two inteins at different points in the primary sequence, thereby allowing the two desired trans-splicing reactions to occur simultaneously with high fidelity and efficiency. Indeed, an elegant feature of this approach is the self-assembly of the four intein fragments which not only allows the one-pot synthesis of the product, but significantly reduces the concentration dependence of this third-order reaction. Potential drawbacks of the strategy are that it results in the insertion of several additional amino acids at the two splice junctions and its requirement for relatively harsh reaction conditions (both chemical denaturants and elevated temperatures are at various points required).

Although EPL has yet to be used for the selective isotopic labeling of an internal region of a protein, the biosynthetic tools required for this purpose have been reported.[28,29] These make use of a so-called sequential EPL reaction which allows a synthetic peptide or a recombinant polypeptide to be inserted into the interior of a recombinant protein framework (Fig. 2B). In this strategy, the polypeptide insert contains both an α-thioester and a cryptic α-Cys masked by a factor Xa cleavable pro-sequence. This reversible cysteine protection is necessary to prevent the insert reacting with itself in either an inter- or intramolecular fashion.[24,27] In the first step, the insert is reacted through its α-thioester group to the α-Cys moiety of a C-terminal recombinant polypeptide. The pro-sequence is then removed from this intermediate ligation product by treatment with factor Xa, to reveal the requisite α-Cys for the second ligation reaction, this time with the N-terminal polypeptide α-thioester segment. In this way, the regioselective insertion is achieved. As with all EPL reactions, the chemistry can be performed in the presence or absence of chaotropic reagents and procedures are now available which allow all of the steps to be performed either in solution[28] or on the solid-phase.[29]

In principle, sequential EPL permits the segmental isotopic labeling of internal regions of proteins. In the present work, the feasibility of this is demonstrated through the synthesis of the eukaryotic adaptor protein, c-Crk-II, from three recombinant polypeptide fragments. Moreover, we show that the resulting multi-component ligation product is fully biologically active in a series of ligand binding studies. Thus, sequential EPL permits both the *endo* and *exo* segmental isotopic labeling of this protein.

Results and Discussion

Synthetic design

c-Crk is a member of the so-called 'adaptor protein' family of intracellular signaling proteins. The major role of adaptor proteins such as Grb2, Nck, and Crk, is to recruit proline-rich effector molecules to tyrosine phosphorylated kinases or their substrates, a function which they mediate through their Src homology 3 (SH3) and SH2 domains.[40,41] Thus, adaptor proteins play a critical role in regulating signaling pathways, and c-Crk has been implicated in several such processes including integrin signaling and growth factor signaling.[40–42] Two cellular forms of Crk have been identified, c-Crk-II and c-Crk-L, both of which have the domain arrangement, SH2-SH3-SH3. In addition, both contain a conserved tyrosine (located between the two SH3 domains) which upon phosphorylation by the non-receptor protein tyrosine kinase, c-Abl, is thought to regulate c-Crk-II function by inducing an intramolecular association with the SH2 domain.[43,44] Interestingly, alternative splicing of the *crk-II* gene leads to a truncated protein (known as c-Crk-I) which lacks the C-terminal SH3 domain and the key tyrosine, and which is more transforming than c-Crk-II in certain cultured cell lines.[40]

Although several proteins have been shown to interact with the SH2 and central SH3 domains of c-Crk, no proteins that bind to the C-terminal SH3 domain have been identified and the function of this domain in signaling is unknown. Fluorescence resonance energy transfer (FRET) measurements recently carried in our laboratory,[29] indicate that the c-Crk-II protein has a compacted domain architecture in solution in the absence of phosphorylation. This data suggests that distal regions of the protein interact with one another in an intramolecular fashion. In order to fully explore this possibility at the molecular level, it is important to have high resolution structural information on c-Crk-II. Unfortunately, the full length protein has resisted X-ray crystallographic analysis, and at ~33 kDa is on the upper limit of detailed solution structure analysis by NMR.[45] Consequently, we believe this structural problem is well suited to the segmental isotopic labeling approach described earlier.

As illustrated in Fig. 3, our objective is to assembly the full length mouse c-Crk-II protein from three recombinant polypeptide fragments (each encompassing one of the SH domains), thereby allowing the systematic segmental isotopic labeling of the protein. The two ligation sites G^{124}–S^{125} and G^{207}–S^{208} are both located in the linker regions between the SH domains, and in each case a conservative Ser→Cys mutation is required to facilitate the ligation reaction. Since these mutations are in the linker regions, they are unlikely to affect the folding properties of the assembled protein. It is also important to note that the N-terminal and middle fragments will have a glycine

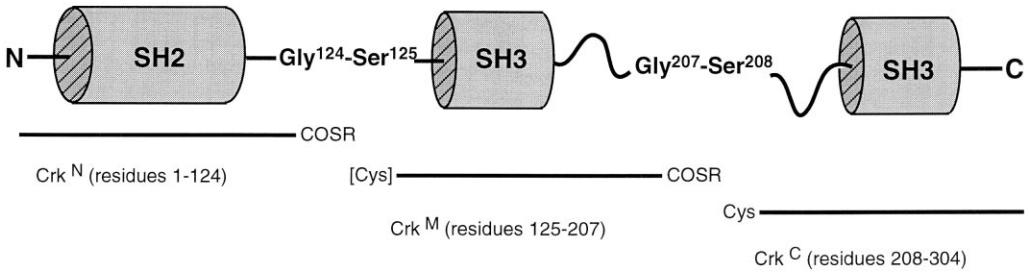

Figure 3. The domain structure of the 304 residue mouse c-Crk-II adaptor protein. The protein contains 3 Src homology domains (SH2, residues [13–118]; central-SH3, residues [132–193]; C-terminal SH3, residues [242–304]) connected by linkers of different length. The ligation sites, G^{124}–S^{125} and G^{207}–S^{208} choosen for the assembly of Crk-II from three individually expressed protein domains (Crk^N, Crk^M and Crk^C) are both located in the linker regions, and in each case a Ser→Cys mutation is required to facilitate the ligation chemistry.

residue at their C-terminus, a feature which is predicted to result in minimal in vivo cleavage of the corresponding intein-fusion proteins[21,46] and, in addition, to promote high yields of thiol induced cleavage and expressed protein ligation.[23]

Expression and purification of fragments

All three c-Crk-II fragments (hereafter designated Crk^N [residues 1–124 and containing the SH2 domain], Crk^M [residues 125–207 and containing the central SH3 domain] and Crk^C [residues 208–304 and containing the C-terminal SH3 domain]) were expressed as C-terminal intein fusions. In addition, an N-terminal factor Xa removable pro sequence (referred to as Xa) was appended to both the Crk^M and Crk^C fragments in order to facilitate controlled sequential EPL.[28,29] Approximately one hundred different intein domains have been identified to date which, broadly speaking, can be divided into two groups; those which contain a core homing endonuclease domain (which is not involved in protein splicing) and those which do not. The latter are in the minority and are often referred to as mini-inteins.[17] *E.coli* expression vectors are now commercially available which allow the generation of protein-fusions to two different engineered inteins, the *Saccharomyces cerevisiae* VMA intein (454 residues and containing an endonuclease domain) and the *Mycobacterium xenopi* GyrA mini-intein (198 residues). We have found that the final yield of the desired recombinant protein α-thioester can depend strongly on which intein-fusion system is employed.[47] The two main contributing factors to low yield are in vivo cleavage of the fusion protein and low levels of soluble expression (note that it in our experience, soluble expression of the intein-fusion is desirable due to the poor yields associated with reconstituting a functional intein from bacterial inclusion bodies). In practical terms it is difficult to know, a priori, which intein expression system will work best for a given polypeptide and thus it is advisable to try both in parallel.

Two sets of bacterial expression vectors were prepared using standard PCR subcloning methods. In one set, the appropriate fragment of the c-Crk-II gene (i.e. Crk^N, Crk^M or Crk^C) was cloned in-frame into a pTYB vector which, upon expression, generates the corresponding VMA-intein-CBD fusion protein (where CBD is a chitin binding domain affinity handle). The Crk^N and Crk^C gene fragments were also cloned into a pTXB vector which generates the corresponding GyrA-intein-CBD fusion protein (note, for technical reasons the Crk^M fragment could not be cloned into the pTXB system). The various expression vectors were then transformed into BL21 cells and a series of preliminary studies undertaken to identify the optimal intein system and expression conditions (cell density, IPTG concentration and induction time/temperature were varied) for the three Crk fragments. The highest levels of soluble expression of both the Crk^N and Crk^C fragments were observed using the GyrA intein at an induction temperature of 37°C, whereas the Crk^M fragment in the VMA intein was found to express best at 25°C. Thus, large scale over-expression of the three fusion proteins, Crk^N-GyrA-CBD, Xa-Crk^M-VMA-CBD, Xa-Crk^C-GyrA-CBD, was undertaken.

Xa-Crk^C-GyrA-CBD was affinity purified from bacterial cell lysate using chitin beads. These beads were then treated overnight with DTT which released the Xa-Crk^C fragment as an α-COOH. Following dialysis to remove the reducing agent, Xa-Crk^C was digested with factor Xa to unmask the N-terminal cysteine necessary for the ligation reaction. The yield of this proteolytic deprotection step was ~70% after overnight digestion as determined by SDS-PAGE. An initial attempt to purify the digestion product, α-Cys-Crk^C, from starting material and factor Xa using gel filtration (Superdex 75 column) proved unsuccessful. (Note, efficient removal of the protease is particularly important since it could interfere with subsequent steps, in contrast, complete removal of the unreactive starting material, Xa-Crk^C, is less critical at this stage.) Preparative scale reversed-phase HPLC was, however, found to have sufficient resolving power to remove the protease from α-Cys-Crk^C, although small amounts of undigested starting material still remained. Using the above procedure it was possible to obtain ~12 mg of this semi-pure α-Cys-Crk^C material per L of bacterial culture (Fig. 4A, lane 1).

The Xa-Crk^M-VMA-CBD fusion protein was found to undergo appreciable in vivo cleavage between the Crk^M and VMA domains as judged by SDS-PAGE analysis of the supernatants. This side-reaction could be minimized to ~30% through optimization of the expression and purification protocols. It is intriguing that in vivo cleavage was observed for Xa-Crk^M-VMA-CBD but not Xa-Crk^C-GyrA-CBD - both proteins have the general arrangement Xa-Cys-linker-SH3-intein and it might be expected that the Cys thiol could react in an intramolecular fashion with the thioester connecting the SH3 domain to the intein to induce cleavage

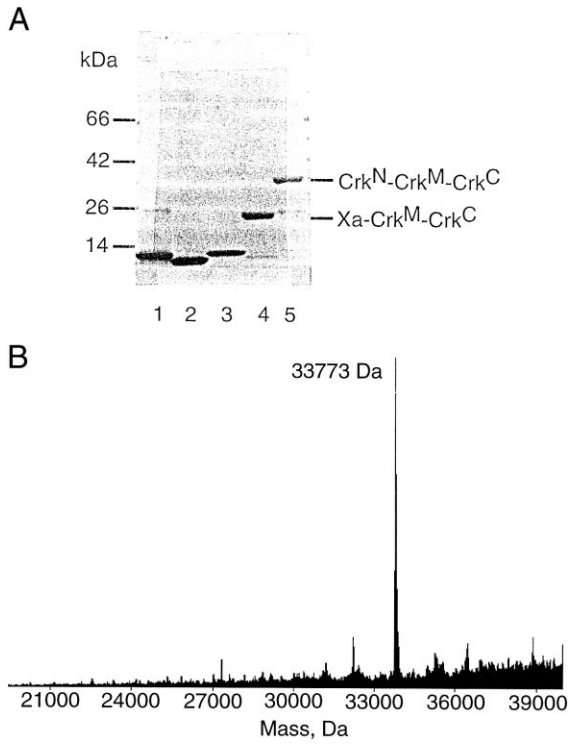

Figure 4. Sequential expressed protein ligation of the three c-Crk-II fragments, Crk^N, Crk^M and Crk^C. (A) Coomassie stained SDS-PAGE (4–20% gel) analysis of: Lane 1, α-Cys-Crk^C; Lane 2, Xa-Crk^M ethyl-α-thioester; Lane 3, Crk^N ethyl-α-thioester; Lane 4, Xa-Crk^M-Crk^C; Lane 5, fully assembled Crk^N-Crk^M-Crk^C, (B) Reconstructed electrospray mass spectrum of the assembled c-Crk-II protein (expected mass=33,772.6 Da).

(i.e. macrothiolactonization). The major difference between the fusion proteins is the length of the linker region between the Cys and the beginning of the SH3 domain, 7 residues in the case of the Xa-Crk^MVMA-CBD system compared with 34 residues for Xa-Crk^C-GyrA-CBD. We speculate that the shorter linker in the Xa-Crk^M-VMA-CBD system, in combination with the topology of the SH3 domain fold (the N- and C-termini are juxtaposed [48]) leads to higher levels of in vivo cleavage due to more facile macrothiolactonization. Consistent with this hypothesis, treatment of affinity immobilized Xa-Crk^M-VMA-CBD with ethanethiol gave rise to two cleavage products which could be resolved by analytical HPLC (data not shown). One species had a mass corresponding to the expected ethyl α-thioester derivative of Xa-Crk^M, whereas the other species had a mass corresponding to the macrothiolactone of Xa-Crk^M (the ratio of the components depended on the amount of ethanethiol present). Base hydrolysis of both products generated the α-COOH derivative of Xa-Crk^M based on HPLC and electrospray mass spectrometry (ESMS). Since both cleavage products contain a reactive α-thioester moiety and are interconvertible through transthioesterification, they were combined after HPLC purification of the thiol cleavage mixture. Using this protocol, the overall yield of purified Xa-Crk^M α-thioester was 8 mg/L of cultured cells (Fig. 4A, lane 2).

The Crk^N-GyrA-CBD fusion protein was affinity purified from cell supernatants using chitin beads. Treatment of these beads with ethanethiol yielded the desired ethyl α-thioester derivative of Crk^N in excellent yield (11 mg/L of cultured cells) and high purity (Fig. 4A, lane 3). This cleavage supernatant was used directly in the ligation reactions described below.

Sequential EPL in solution

The overall reaction scheme followed is essentially that outlined in Fig. 2B. In the first step, purified Xa-Crk^M α-thioester and α-Cys-Crk^C were combined in ligation buffer (0.2 M phosphate, pH 7.2, 0.2 M NaCl) in the presence of 2% MESNA (2% w/v) and ethanethiol (1% v/v). Inclusion of a small amount of chaotrope (0.5 M GuHCl) in this buffer permitted high concentrations of the two reactants to be achieved (1.8 mM and 1.0 mM, respectively) thereby improving the kinetics of the ligation reaction. Based on analytical HPLC and ESMS analysis of the ligation mixture, the reaction was terminated after 48 h and the desired ligation product Xa-Crk^M-Crk^C purified by preparative HPLC in 40% recovered yield (Fig. 4A, lane 4). The modest yield of the purified product is on the lower end of what is considered typical for these reactions[12] and was primarily attributable to competing hydrolysis of the Xa-Crk^M α-thioester fragment. Inclusion of excess thiols in the ligation soup usually minimizes α-thioester hydrolysis[12], however, this side-reaction is significant in the case of the Xa-Crk^M α-thioester segment for reasons which are unknown. We are currently investigating whether this phenomenon is related to the tendency of Xa-Crk^M α-thioester to undergo macrothiolactonization.

In the second step, Xa-Crk^M-Crk^C was treated with factor Xa for 4 h at 20°C in order to give α-Cys-Crk^M-Crk^C which contains the N-terminal Cys required for the next ligation reaction. This proteolytic deprotection was found to be extremely efficient as determined by HPLC and ESMS analysis of the digestion mixture (no starting material could be detected). The crude digestion mixture was then added directly to freshly generated Crk^N ethyl α-thioester in order to initiate the third and final step of the sequential EPL synthesis. After 18 h, analytical HPLC and ESMS analysis indicated the ligation reaction had essentially gone to completion (>90% based on peak integration in analytical HPLC). The desired ligation product Crk^N-Crk^M-Crk^C (i.e. c-Crk-II) was isolated using HPLC in satisfactory yield (0.65 mg, 0.02 μmol, 50%) (Fig. 4A, lane 5 and Fig. 4B).

Biochemical activity of ligation product

The final ligation product, Crk^N-Crk^M-Crk^C, contains two Ser→Cys mutations and was subjected to a final HPLC purification step. Thus, it was important to assess whether the ligation product retained the expected biological activities of the c-Crk-II protein. As mentioned previously, the SH2 and central SH3 domains of the protein are known to interact with a variety of proteins and high affinity peptide ligands for these two domains have been identified (no ligands are available for the C-terminal SH3 domain). Consequently, we determined whether these two Src homology domains within the refolded Crk^N-Crk^M-Crk^C had the expected ligand binding properties.

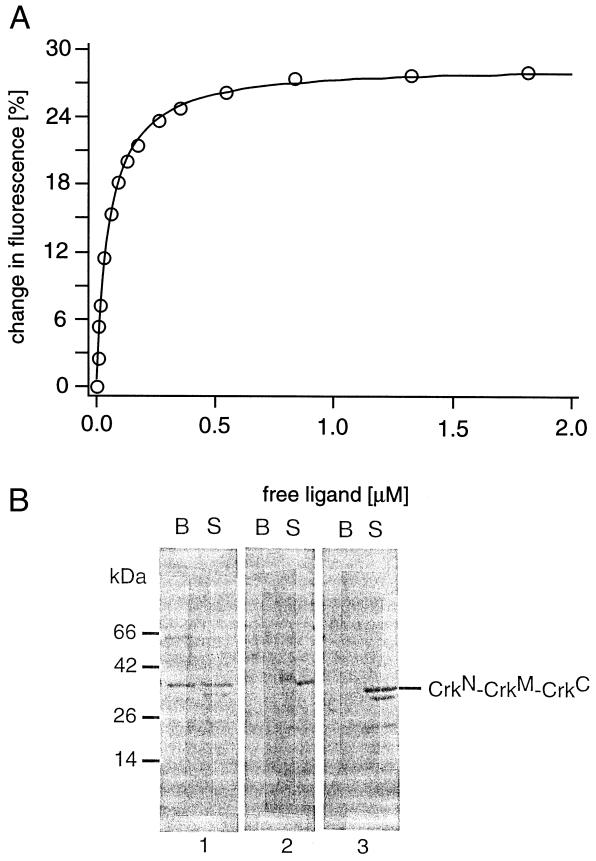

Figure 5. Ligand binding properties of the assembled c-Crk-II protein. (A) Fluorescence-based ligand binding assay. The high affinity ligand (PPPALPPKRRR-NH_2)[49] for the central SH3 domain was titrated into a solution of Crk^N-Crk^M-Crk^C. Excitation was at 290 nm and the fluorescence emission was monitored at 345 nm. (B) Affinity pull-down assay probing the activity of the SH2 domain. The phosphopeptide ligand for the SH2 domain (Cys-Ahx-Gly-pTyr-Ala-Gln-Pro-Ser-NH_2) was covalently attached to agarose beads. SDS-PAGE (4–20% gel) analysis of the beads (B) and supernatant (S) is shown after: (1) Incubation of Crk^N-Crk^M-Crk^C with the peptide-affinity beads; (2) Elution of Crk^N-Crk^M-Crk^C from the immobilized phosphopeptide by addition of soluble phosphopeptide; (3) Incubation of Crk^N-Crk^M-Crk^C with the agarose beads alone.

A fluorescence-based titration assay was used to determine the affinity of a proline-rich peptide ligand, PPPALPPKRRR-NH_2,[49] for the central SH3 domain (Fig. 5A). The measured dissociation constant (K_D) of 53 ± 3 nM was very similar to that previously determined for the wild-type SH3 domain ((K_D=118 ± 10 nM).[49] To evaluate the function of the SH2 domain within the reassembled Crk-II protein, an affinity pull-down assay was performed in which binding to an immobilized phosphopeptide ligand Cys-Ahx-Gly-Pro-phosphoTyr-Ala-Gln-Pro-Ser-NH_2 (based on the known phosphorylation motif in c-Crk-II) was probed (Fig. 5B). Refolded Crk^N-Crk^M-Crk^C was found to bind efficiently to the peptide-affinity column but not to the affinity matrix alone. Moreover, the immobilized Crk^N-Crk^M-Crk^C could be competitively eluted from the column using excess soluble phosphopeptide, which is consistent with a specific SH2-peptide interaction.

Taken together, these ligand binding studies indicate that the SH2 and central SH3 domains within the final ligation product are biochemically active and thus properly folded. These studies do not, however, provide any information on the folded state of the C-terminal SH3 domain, nor do they yield any information on whether the interdomain architecture of the ligation product is the same as the wild-type protein. Future biochemical and structural studies will address these issues in more detail.

Summary

By allowing the controlled assembly of 3 or more polypeptide segments, sequential EPL permits the segmental isotopic labeling of internal regions of large proteins. In the present work we have demonstrated that it is possible to assemble the 304 amino acid protein, c-Crk-II, from three recombinant polypeptide segments and that, to the first approximation, the purified reassembled protein retains the ligand binding properties of the wild-type protein. Optimization of the protein expression and purification steps make it possible to prepare tens of milligram amounts of the three reactive protein segments; Crk^N ethyl α-thioester, Xa-Crk^M α-thioester and α-Cys-Crk^C. This coupled with the moderate-to-high yields associated with the sequential EPL process, suggests that it will be practical to prepare NMR sample quantities of the segmental labeled version of the protein in which any of the three segments is labeled with NMR probe nuclei.

Experimental

General materials and methods

Analytical gradient HPLC was performed on a Hewlett-Packard 1100 series instrument with 214 and 280 nm detection using a Vydac C18 column (5 micron, 4.6×150 mm) at a flow rate of 1 mL/min. Preparative HPLC was performed on a Waters DeltaPrep 4000 system fitted with a Waters 486 tunable absorbance detector using a Vydac C18 column (15–20 micron, 50×250 mm) at a flow rate of 30 mL/min. All runs used linear gradients of 0.1% aqueous TFA (solvent A) vs 90% acetonitrile plus 0.1% TFA (solvent B). Electrospray mass spectrometric (ESMS) analysis was routinely applied to all synthetic peptides and components of reaction mixtures. ESMS was performed on a Sciex API-100 single quadrupole electrospray mass spectrometer. Calculated masses were obtained using the program MacProMass (Sunil Vemuri and Terry Lee, City of Hope, Duarte, CA) and reflect average isotope compositions. Expressed proteins were routinely analyzed on SDS-PAGE using standard procedures.

Construction of plasmids

The DNA encoding the Crk^N fragment (residues Met-1 to Gly-124) of murine c-Crk was amplified by PCR from a full length clone of m-Crk using the oligonucleotide primers, mCrk#1 (5′-ATG AAA GCG CGG AGG CTC ATA TGG CGG GCA ACT TCG ACT CG-3′) and mCrk#2 (5′-CTG CCT GAG AGC TCT TCC GCA ACC CTG CCT TGA TCT GGC-3′). The primers were designed to introduce a

NdeI site (mCrk#1) on the 5′-end and a SapI site (mCrk#2) on the 3′-end of the gene fragment. The PCR product was purified and digested with NdeI and SapI and then recloned into the NdeI and SapI treated pTXB1 (GyrA intein) and pTYB1 (VMA intein) plasmids (both New England Biolabs).

The DNA encoding the Crk^C fragment (residues His-209 to Ser-304) was amplified by PCR from an m-Crk clone using the oligonucleotide primers, mCrk#3 (5′-AAA AAA AAA AAA AAA CAT ATG CTG ATC GAA GGT CGT TGT CAC CCA CAG CCA CTG GGT GG-3′) and mCrk#4 (5′-GCA AAC TGG CTC TTC CGC AGC CGC TGA AGT CCT CAT CGG G -3′). The primers were designed to introduce a NdeI site (mCrk#3) on the 5′-end and a SapI site (mCrk#4) on the 3′-end of the gene fragment. In addition, the mCrk#3 primer was designed to introduce the sequence MLIEGRC (referred to as Xa) immediately N-terminal to residue His-209. The PCR product was purified and digested with NdeI and SapI and then recloned into the NdeI and SapI treated pTXB1 (GyrA intein) or pTYB1 (VMA intein) plasmids.

The cDNA encoding the Crk^M fragment (residues Gly-126 to Glu-206) was amplified by PCR from a full length clone of m-Crk using the oligonucleotide primers, mCrk#5 (5′-AAA AAA AAA AAA AAA ACA TAT GCT GAT CGA AGG TCG TTG TGG AGT GAT TCT CAG GCA GG-3′) and mCrk#6 (5′-AAA AAA AAA GTG GGA CCG CTC CTG GTT ACC TCC AAT CAG AGC-3′). The mCrk#5 primer was designed to introduce a NdeI site on the 5′ end and to introduce the sequence MLIEGRC immediately *N*-terminal to residue Gly-126. The mCrk#6 primer was designed to introduce a BsrBI site on the 3′ end. After purification, the PCR product was digested with BsrBI and NdeI and cloned into a NdeI and SmaI treated pTYB2 plasmid (New England Biolabs). The pTYB2 plasmid introduces a Gly at the C-terminus of the expressed fragment which corresponds to Gly-207 of c-Crk-II

The resulting T7-driven expression plasmids were shown to be free of mutations by automated DNA sequencing.

Overexpression of Crk^N, Crk^M and Crk^C. Optimal expression vectors and conditions were determined for each of the three fragments using standard screening protocols (data not shown).

E.coli BL21 cells transformed with the plasmids $pTXB1_{CrkN}$, $pTYB2_{CrkM}$ or $pTXB1_{CrkC}$ were grown at 37°C in Luria Bertani (LB) medium containing 100 μg/ml ampicillin to an OD_{590} of 0.4–0.5. $pTXB1_{CrkN}$, and $pTXB1_{CrkC}$ transformed cells were induced at 37°C with 0.5 mM IPTG for 5 h to give soluble expression of the desired intein fusion protein. Cells transformed with the $pTYB2_{CrkM}$ plasmid were induced at 25°C with 0.5 mM IPTG for 4 h. Cells were harvested by centrifugation (8500×g, 4°C, 5 min) and stored as a pellet at −80°C. After thawing, cells were resuspended in 25 mL of lysis buffer (20 mM phosphate, pH 7.0, 500 mM NaCl, 1 mM EDTA, 5% glycerol) and disrupted using a French press. After centrifugation (31000×g, 4°C, 30 min), the supernatants were diluted with 75 mL of column buffer (20 mM HEPES, pH 7.0, 250 mM NaCl, 1 mM EDTA, 0.1% Triton X-100) and then loaded onto chitin beads (5 mL/L of cultured cells) equilibrated in column buffer. The resulting affinity beads were then thoroughly washed with column buffer.

Preparation of α-Cys-Crk^C. The chitin beads carrying the immobilized Xa-Crk^C-GyrA-CBD fusion protein were equilibrated with ligation buffer (200 mM phosphate, pH 7.2, 200 mM NaCl) and the Xa-Crk^C fragment was cleaved from the beads by treatment with ligation buffer containing 50 mM DTT for 24 h. The beads were filtered off and washed with ligation buffer. The combined supernatants were then dialysed extensively against 50 mM Tris HCl, pH 8.0, 100 mM NaCl, 1 mM $CaCl_2$. Digestion of Xa-Crk^C with factor Xa (25 U of enzyme/mg Xa-Crk^C) was performed in Tris buffer for 2 h at 20°C after which the mixture was purified by reversed phase HPLC (31–38% buffer B over 60 min). Note that undigested starting material, Xa-Crk^C, could not be completely resolved from the digestion product, α-Cys-Crk^C. This HPLC purified material thus contained a mixture of Xa-Crk^C (30%, ESMS; observed=11,444.3±2.0 Da, expected for $C_{508}H_{793}N_{145}O_{151}S_3$=11,443.9 Da) and α-Cys-$Crk^C$ (70%, ESMS; observed=10,743.3±0.8 Da, calculated mass for $C_{478}H_{740}N_{136}O_{143}S_2$=10744.1 Da). The overall yield of this semi-pure α-Cys-Crk^C was 12 mg/L cultured cells.

Preparation of Xa-Crk^M ethyl-α-thioester. Chitin beads loaded with the Xa-Crk^M-VMA-CBD fusion protein were equilibrated with ligation buffer containing 2% (v/v) EtSH and incubated overnight at 20°C. The protein was purified from the supernatant by HPLC (30–50% buffer B over 60 min) and lyophilized. In addition to the expected Xa-Crk^M ethyl-α-thioester product (ESMS; observed= 10,173.9±2.6 Da, calculated mass for=10,173.3 Da), a second species was isolated which corresponded to the thiolactone derivative (ESMS; observed=10,111.6±2.2 Da, calculated mass for $C_{446}H_{692}N_{124}O_{139}S_3$=10,111.3 Da). These two materials were combined to give a final yield of 8 mg/L of cultured cells.

Preparation of Crk^N ethyl-α-thioester. Chitin beads loaded with the Crk^N-GyrA-CBD fusion protein were treated with ligation buffer containing 2% EtSH at 20°C. After 24 h the supernatant containing the Crk^N ethyl-α-thioester (ESMS; observed=13,678.9±2.5 Da, calculated mass for $C_{608}H_{940}N_{174}O_{185}S$=13,679.1 Da) at a concentration of 12 mg/mL was removed and used in the ligation reaction without further purification. The overall yield of Crk^N ethyl-α-thioester was ~11 mg/L cultured cells.

Ligation of Xa-Crk^M ethyl-α-thioester and α-Cys-Crk^C. α-Cys-Crk^C (2 mg, 0.175 μmol) was dissolved in 100 μL of ligation buffer containing 0.5 M GuHCl, 2% (w/v) MESNA and 1% (v/v) EtSH. To this solution was added lyophilized Xa-Crk^M ethyl-α-thioester (1 mg, 0.100 μmol). After incubation for 48 h at room temperature the expected ligation product Xa-Crk^M-Crk^C (ESMS; observed= 20,855.2±5.5 Da, calculated mass for $C_{924}H_{1432}N_{260}O_{282}S_5$=20,855.4 Da) was purified by HPLC (31–48% buffer B over 60 min) in 40% yield (0.8 mg, 0.04 μmol).

Digestion of Xa-Crk^{M}-Crk^{C} with factor Xa protease. Purified Xa-Crk^{M}-Crk^{C} (0.8 mg, 0.04 μmol) was dissolved in 100 μL of ligation buffer and 40 U of factor Xa in 40 μL of H_2O added. The digestion was allowed to proceed for 4 h at 20°C to give α-Cys-Crk^{M}-Crk^{C} (ESMS; observed=20,158.2±1.7 Da, calculated mass for $C_{894}H_{1379}N_{251}O_{274}S_4$=20,155.5 Da) in high yield (>90%) as indicated by analytical HPLC and ESMS. This digestion mixture was used directly in the subsequent ligation step.

Ligation of α-Cys-Crk^{M}-Crk^{C} and Crk^{N} ethyl-α-thioester. The digestion mixture from above was mixed with a freshly generated solution of Crk^{N} ethyl-α-thioester (300 μL at 12 mg/ml) in ligation buffer containing 3% (w/v) MESNA and 2% (v/v) EtSH. After incubation for 18 h at 20°C, the expected ligation product Crk^{N}-Crk^{M}-Crk^{C} (ESMS; observed=33,773.1±5.6 Da, calculated mass for $C_{1500}H_{2313}N_{425}O_{459}S_4$ =33,772.6 Da) was purified by HPLC (31–50% buffer B over 60 min) in 50% recovered yield (0.65 mg, 0.02 μmol).

Fluorescence-based ligand binding assay. A solution of ligation product, Crk^{N}-Crk^{M}-Crk^{C} (0.180 μM) in a buffer containing 2 mM DTT, 150 mM NaCl, 10 mM sodium phosphate buffer pH 7.3 was titrated with increasing concentrations of a high affinity ligand for the central SH3 domain of Crk, PPPALPPKRRR-NH_2[49], and the change in the intrinsic fluorescence (predominantly tryptophan) of the protein monitored as a function of ligand concentration. Experiments were conducted at 18°C in a stirred 1 cm-path-length cell using a SPEX FL3-11C fluorimeter. Excitation was at 290 nm with a 3 nm slit and the fluorescence emission was monitored at 345 nm through a 6 nm slit. The dissociation constant was determined by changes in the fluorescence of the protein solution upon addition of the corresponding peptide ligand at defined concentrations; calculations were made assuming formation of a 1:1 complex.[50]

Affinity pull-down assay. Crk^{N}-Crk^{M}-Crk^{C} (~4 μg) was dissolved in 50 μL of a loading buffer (50 mM phosphate, pH 7.2, 150 mM NaCl) and incubated overnight with 10 μL Sulfolink agarose beads (Pierce) to which the SH2-binding peptide, Cys-Ahx-Gly-Pro-pTyr-Ala-Gln-Pro-Ser-NH_2, had been covalently attached (1000 pmole peptide/μL of beads) as per manufacturer's protocol. As a control, a similar experiment was performed using the agarose affinity matrix capped with mercaptoethanol. Competition experiments were performed by eluting the loaded affinity beads with a solution containing 5 mM of soluble phosphopeptide in loading buffer.

Acknowledgements

We thank Drs D. Cowburn and R. Birge (Rockefeller University) for many useful discussions. This research was supported by the National Institutes of Health (GM55843, T. W. M.), the Burroughs-Wellcome fund (G. J. C.) and a Merck postdoctoral fellowship (U. K. B.).

References

1. Muir, T. W.; Sondhi, D.; Cole, P. A. *Proc. Natl. Acad. Sci. USA* **1998**, *95*, 6705–6710.
2. Severinov, K.; Muir, T. W. *J. Biol. Chem.* **1998**, *273*, 16205–16209.
3. Evans Jr., T. C.; Benner, J.; Xu, M.-Q. *Protein Sci.* **1998**, *7*, 2256–2264.
4. Cornish, V. W.; Mendel, D.; Schultz, P. G. *Angew. Chem., Int. Ed. Engl.* **1995**, *34*, 621–633.
5. Chen, Y.; Ebright, Y. W.; Ebright, R. H. *Science* **1994**, *265*, 90–92.
6. Wallace, C. J. A. *Curr. Opin. Biotechnol.* **1995**, *6*, 403–410.
7. Wilken, J.; Kent, S. B. H. *Curr. Opin. Biotechnol.* **1998**, *9*, 412–426.
8. Jackson, D. Y.; Burnier, J.; Quan, C.; Stanley, M.; Tom, J.; Wells, J. A. *Science* **1994**, *266*, 243–247.
9. Cotton, G. J.; Muir, T. W. *Chem. Biol.* **1999**, *6*, R247–R256.
10. Hruby, V. J.; Al-Obeidi, F.; Kazmierski, W. *Biochem. J.* **1990**, *268*, 249–262.
11. Dawson, P. E.; Muir, T. W.; Clark-Lewis, I.; Kent, S. B. H. *Science* **1994**, *266*, 776–779.
12. Muir, T. W.; Dawson, P. E.; Kent, S. B. H. *Methods Enzymol.* **1997**, *289*, 266–298.
13. Perler, F. B. *Cell* **1998**, *92*, 1–4.
14. Xu, M. -Q.; Perler, F. B. *EMBO J.* **1996**, *15*, 5146–5153.
15. Paulus, H. *Chem. Soc. Rev.* **1998**, *27*, 375–386.
16. Noren, C. J.; Wang, J.; Perler, F. B. *Angew. Chem., Int. Ed. Engl.* **2000**, *39*, 450–466.
17. Perler, F. B. *Nucleic Acids Res.* **2000**, *28*, 344–345.
18. Erlanson, D. A.; Chytil, M.; Verdine, G. L. *Chem. Biol.* **1996**, *3*, 981–991.
19. Evans, T. C.; Benner, J.; Xu, M. Q. *J. Biol. Chem.* **1999**, *274*, 3923–3926.
20. Mathys, S.; Evans Jr., T. C.; Chute, C. I.; Wu, H.; Chong, S.; Benner, J.; Liu, X. Q.; Xu, M.-Q. *Gene* **1999**, *231*, 1–13.
21. Southworth, M. W.; Amaya, K.; Evans, T. C.; Xu, M.-Q.; Perler, F. B. *BioTechniques* **1999**, *27*, 110–120.
22. Roy, R. S.; Allen, O.; Walsh, C. T. *Chem. Biol.* **1999**, *6*, 789–799.
23. Ayers, B.; Blaschke, U. K.; Camarero, J. A.; Cotton, G. J.; Holford, M.; Muir, T. W. *Biopolymers* **1999**, *51*, 343–354.
24. Camarero, J. A.; Muir, T. W. *J. Am. Chem. Soc.* **1999**, *121*, 5597–5598.
25. Iwai, H.; Pluckthun, A. *FEBS Lett.* **1999**, *459*, 166–172.
26. Scott, C. P.; AbelSantos, E.; Wall, M.; Wahnon, D. C.; Benkovic, S. J. *Proc. Natl. Acad. Sci. USA* **1999**, *96*, 13638–13643.
27. Evans Jr, T. C.; Benner, J.; Xu, M.-Q. *J. Biol. Chem.* **1999**, *274*, 18359–18363.
28. Cotton, G.; Ayers, B.; Xu, R.; Muir, T. W. *J. Am. Chem. Soc.* **1999**, *121*, 1100–1101.
29. Cotton, G. J.; Muir, T. W. *Chem. Biol.* **2000**, *7*, 253–261.
30. Yamazaki, T.; Otomo, T.; Oda, N.; Kyogoku, Y.; Uegaki, K.; Ito, N.; Ishino, Y.; Nakamura, H. *J. Am. Chem. Soc.* **1998**, *120*, 3391–3392.
31. Xu, R.; Ayers, B.; Cowburn, D.; Muir, T. W. *Proc. Natl. Acad. Sci. USA* **1999**, *96*, 388–393.
32. Otomo, T.; Teruya, K.; Uegaki, K.; Yamazaki, T.; Kyogoku, Y. *J. Biomol. NMR* **1999**, *14*, 105–114.
33. Otomo, T.; Ito, N.; Kyogoku, Y.; Yamazaki, T. *Biochemistry* **1999**, *38*, 16040–16044.

34. Pervushin, K.; Riek, R.; Wider, G.; Wuthrich, K. *Proc. Natl. Acad. Sci. USA* **1997**, *94*, 12366–12371.
35. Tjandra, N.; Bax, A. *Science* **1997**, 1111–1114.
36. Mills, K. V.; Lew, B. M.; Jiang, S.-Q.; Paulus, H. *Proc. Natl. Acad. Sci. USA* **1998**, *95*, 3543–3548.
37. Wu, H.; Xu, M.-Q.; Liu, X.-Q. *Biochim. Biophys. Acta* **1998**, *1387*, 422–432.
38. Southworth, M. W.; Adam, E.; Panne, D.; Byer, R.; Kautz, R.; Perler, F. B. *EMBO J.* **1998**, *17*, 918–926.
39. Wu, H.; Hu, Z.; Liu, X.-Q. *Proc. Natl. Acad. Sci. USA* **1998**, *95*, 9226–9231.
40. Buday, L. *Biochim. Biophys. Acta* **1999**, *1422*, 187–204.
41. Birge, R. B.; Knudsen, B. S.; Besser, D.; Hanafusa, H. *Genes to Cells* **1996**, *1*, 595–613.
42. Hashimoto, Y.; Katamaya, H.; Kiyokawa, E.; Ota, S.; Kurata, T.; Gotoh, N.; Otsuka, N.; Shibata, M.; Matsuda, M. *J. Biol. Chem.* **1998**, *273*, 17186–17191.
43. Feller, S. M.; Knudsen, B.; Hanafusa, H. *EMBO J.* **1994**, *13*, 2341–2351.
44. Rosen, M. K.; Yamazaki, T.; Gish, G. D.; Kay, C. M.; Pawson, T.; Kay, L. E. *Nature* **1995**, *374*, 477–479.
45. Wider, G.; Wuthrich, K. *Curr. Opin. Struct. Biol.* **1999**, *9*, 594–601.
46. Chong, S.; Williams, K. S.; Wotkowicz, C.; Xu, M.-Q. *J. Biol. Chem.* **1998**, *273*, 10567–10577.
47. Blaschke, U. K.; Siberstein, J.; Muir, T. W. *Methods Enzymol.* **2000**, *328*, 478–496.
48. Wu, X.; Knudsen, B.; Feller, S. M.; Zheng, J.; Sali, A.; Cowburn, D.; Hanafusa, H.; Kuriyan, J. *Structure* **1995**, *3*, 215–226.
49. Posern, G.; Zheng, J.; Knudsen, B. S.; Kardinal, C.; Muller, K. B.; Voss, J.; Shishido, T.; Cowburn, D.; Cheng, G.; Wang, B.; Kruh, G. D.; Burrell, S. K.; Jacobson, C. A.; Lenz, D. M.; Zamborelli, T. J.; Adermann, K.; Hanafusa, H.; Feller, S. M. *Oncogene* **1998**, *16*, 1903–1912.
50. Gosser, Y. Q.; Zheng, J.; Overduin, M.; Mayer, B. J.; Cowburn, D. *Structure* **1995**, *3*, 1075–1086.