Journal of Biomolecular NMR **28:** 357–367, 2004.
KLUWER/ESCOM
© 2004 *Kluwer Academic Publishers. Printed in the Netherlands.*

Using codon optimization, chaperone co-expression, and rational mutagenesis for production and NMR assignments of human eIF2α

Takuhiro Ito & Gerhard Wagner*
Department of Biological Chemistry and Molecular Pharmacology, Harvard Medical School, 240 Longwood Avenue, Boston, MA 02115, U.S.A.

Received 27 August 2003; Accepted 30 October 2003

Key words: chaperone, codon optimization, eIF2α, protein expression, protein structure

Abstract

Producing a well behaved sample at high concentration is one of the main hurdles when starting a new project on an interesting protein. Especially when one attempts to overexpress a eukaryotic protein in bacteria, some difficulties are encountered, such as low expression level, low solubility, or even lack of folded structure. Overexpression in prokaryotic systems is highly desirable for cost-effective production of different isotope-labeled samples needed for NMR studies. Here we describe generally applicable methods for obtaining highly concentrated protein samples efficiently. This approach was developed as we tried to produce a NMR-suitable sample of the 35 kDa human translation initiation factor eIF2α, a protein that expresses poorly in *E. coli* and has very low solubility. First, an *E. coli* codon-optimized gene was synthesized on a thermal cycler, which increased the expression level by a factor of two. Second, we used co-expression of bacterial chaperone proteins, which largely increased the fraction of correctly folded protein found in the soluble phase. Third, we used rational mutagenesis guided by both the sequence alignment among homologues and the homology of one domain to a known fold for improving solubility and stability of the target protein by tenfold. Combining all these methods made it possible to produce from a one-liter preparation a 0.5 mM sample of human eIF2α that showed well-resolved NMR spectra and enabled nearly complete assignment of the protein. These methods may be generally useful for studies of other eukaryotic proteins that are otherwise difficult to express and exhibit poor solubility.

Introduction

Most protein samples used for NMR structure determination are produced in bacterial expression systems, primarily in *Escherichia coli* (*E. coli*). Even though some eukaryotic expression systems have been developed, such as *Pichia pastoris* or *Spodoptera frugiperda* (Bruggert et al., 2003; Morgan et al., 2000), using prokaryotic expression systems is advantageous for labeling with stable isotopes. This is achieved by substituting minimal media with ^{13}C-labeled D-glucose and $^{15}NH_4Cl$ as a sole carbon and nitrogen source, respectively. In addition, larger proteins are commonly perdeuterated by growing bacteria in ^2H_2O and sometimes using ^{2}H-labeled D-glucose in order to reduce relaxation losses due to the efficient ^{1}H-^{13}C dipolar interaction. Recently, several methods for producing perdeuterated proteins containing protonated methyl groups in isoleucine, leucine and valine (ILV) residues have been developed (Gardner and Kay, 1997; Goto et al., 1999; Gross et al., 2003; Hajduk et al., 2000), which are instrumental for solving the global folds of large proteins such as 45 kDa maltose-binding protein (Mueller et al., 2000). However, these labeling methods are expensive since they need special materials, such as [3-^{2}H] ^{13}C α-ketoisovalerate or [3,3-$^2H_2$] ^{13}C α-ketobutyrate, in addition to the regular ^{2}H, ^{13}C, or ^{15}N-labeled material. Furthermore, because the ILV labeling methods are based on the metabolic pathway of the bacteria, it would be hard to use them in the eukaryotic expression systems. Con-

*To whom correspondence should be addressed. E-mail: wagner@hms.harvard.edu sidering all these conditions, it is desirable to have more efficient prokaryotic expression methods for producing eukaryotic proteins that are suitable for NMR studies.

While this goal is obvious, many problems are encountered in attempts to express eukaryotic proteins in bacteria, such as expression in inclusion bodies, incorrect folding, or low expression yield (Kane, 1995; Wall and Pluckthun, 1995). The latter is probably the most severe problem for NMR studies because generally several milligram of the target protein are required for recording the spectra. It is believed that part of the problems arises from the difference of the codon usages between prokaryotes and eukaryotes (Kane, 1995). In other words, products of the eukaryotic gene containing lots of prokaryotic minor codons are supposed to be difficult to overexpress in the prokaryotic system. To overcome this problem, *E. coli* strains that contain an extra plasmid for the minor tRNAs have been developed and are commercially available as *E. coli* Rosetta (Novagen) or *E. coli* Codonplus (Stratagene). Although these strains are efficient for producing large amount of some eukaryotic proteins, other problems are often encountered that cannot be resolved with these strains, such as expression in inclusion bodies, incorrect folding or low solubility of the expressed protein. We encountered problems of low expression yield, incorrect folding and low solubility when pursuing an NMR structure analysis of the human translation initiation factor eIF2α, and standard bacterial expression procedures were not sufficient for pursuing this project.

Here we describe a combination of methods for producing, in an *E. coli* expression system, sufficient amounts of soluble human eIF2α that allowed us to obtain nearly complete resonance assignments and identify its secondary structure. To overcome a difference of codon usages between *E. coli* and eukaryotes, we synthesized a gene optimized for *E. coli* codon usage. To achieve this, a synthetic gene was produced with a simple PCR based method. Having a codon-optimized synthetic gene has significant benefits over employing the Rosetta and Codonplus strains. In the approach reported here, it is possible to add an expression vector for auxiliary folding factors on a pACYC-type plasmid. This produces the chaperone proteins, GroEL-GroES and trigger factor (TF), which prevent the target protein from going into inclusion bodies (Nishihara et al., 2000). This plasmid cannot be expressed in Rosetta and Codonplus strains because they use the same type of plasmid to supply minor tRNAs. In addition, we used a rational strategy to make solubility-enhancing mutants of human eIF2α based on both the sequence alignment with homologues and the homology of one domain with a known fold. The triple mutation, together with short terminal deletions, increased the solubility by a factor of ten, which was sufficient to pursue assignments and structure determination.

Materials and methods

Synthesis of codon-optimized gene

To design a codon-optimized gene, one major codon for each amino acid was selected based on the previous reports of *E. coli* codon usage (GCT for Ala, CGT for Arg, AAC for Asn, GAC for Asp, TGC for Cys, CAG for Gln, GAA for Glu, GGT for Gly, CAG for His, ATC for Ile, CTG for Leu, AAA for Lys, ATG for Met, TTC for Phe, CCG for Pro, TCT for Ser, ACC for Thr, TGG for Trp, TAC for Tyr, and GTT for Val) (Sharp and Li, 1987). Twelve 100-mer DNA oligonucleotides were designed for templates as shown in Figure 1A, and were purchased as 250 nmole scale products with PAGE purification from IDT (Integrated DNA Technologies). All reactions for gene synthesis were performed using Pfu DNA polymerase (Stratagene) on a MiniCycler thermal cycler (MJ Research). Two step reactions were used for the gene synthesis. For the first reaction, all 12 template oligonucleotides were mixed so that the final concentration for each was 0.5 μM, and other required materials (1× native Pfu reaction buffer, 200 μM dNTPs each, and 2 μl Pfu polymerase in 100 μl) were added. The mixture was heated at 94 °C for 3 min, cooled down gently to 55 °C for annealing, and kept at 72 °C for 20 min for elongation. A 2 μl sample from the first reaction product was used as a template for the second PCR-like reaction in a 100 μl mixture, which contained 1× native Pfu reaction buffer, 0.5 μM forward and reverse primers each, 200 μM dNTPs each, and 2 μl Pfu polymerase. The mixture was preheated at 95 °C for 45 s, followed by 35 cycles of heating at 95 °C for 45 s, annealing at 55 °C for 45 s, and elongation at 72 °C for 2.5 min. The final 10-minute elongation reaction was carried out at 72 °C. The amplified product from the second reaction was purified by agarose gel, extracted using Ultrafree-DA (Amicon), and inserted into the pT7Blue vector with the Perfectly Blunt Cloning Kit (Novagen). This created the resulting plasmid, pT7heIF2aopt.

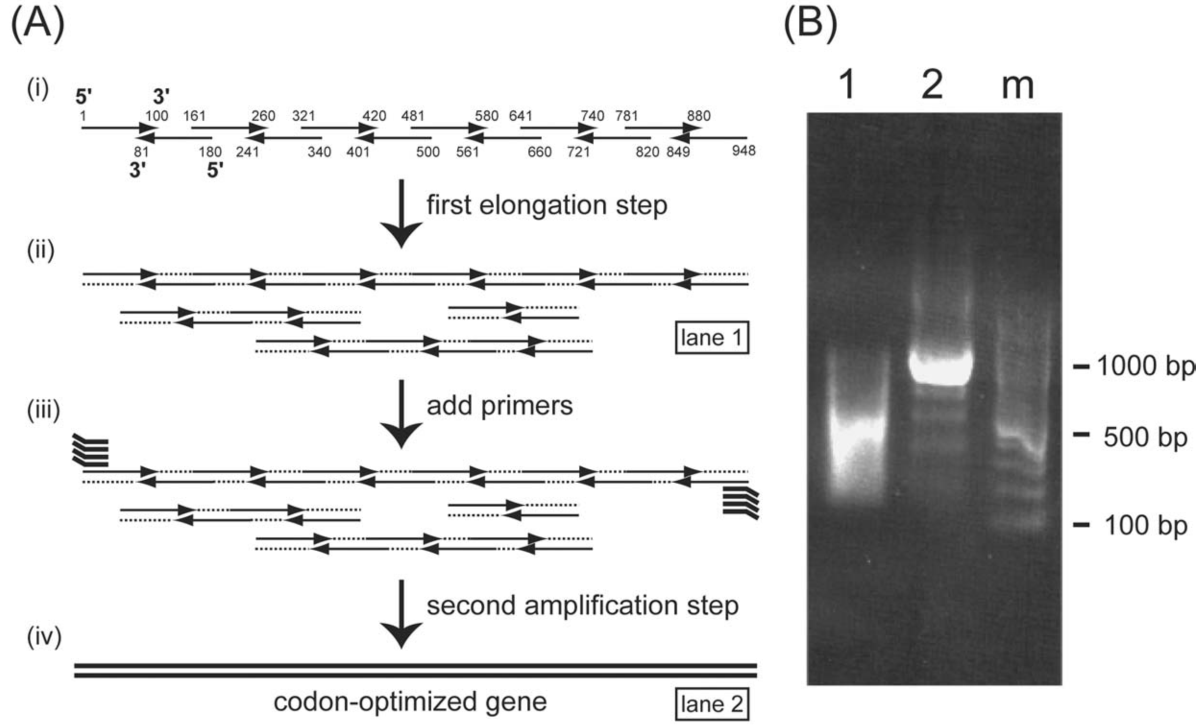

Figure 1. (A) Schematic representation of the synthesis scheme for producing codon-optimized gene. (i) Design of the templates for the gene. Twelve 100mer oligonucleotides were synthesized. Numbers indicate the nucleotide positions at the ends of the oligonucleotides. Arrows indicate the direction of the oligonucleotides, from 5′ to 3′ ends. (ii) Products of the first elongation step. The dotted lines indicate elongated nucleotides. (iii) Second PCR-like amplification step. The primers shown with thick lines were added to amplify the full-length target gene. (iv) Amplified codon-optimized gene. (B) Analysis of the products of the steps. The products of the first elongation step (lane 1) and the second amplification step (lane 2) were analyzed with 1.2% agarose gel electrophoresis with markers in lane m.

Construction of plasmids

Expression plasmids of the native protein were constructed by amplifying target regions in the native cDNA on pGEX-heIF2α or the codon-optimized gene on pT7heIF2aopt by PCR and by inserting them into the pET-30a(+) expression vector (Novagen). The plasmid pGEX-heIF2α was kindly donated by Prof Nahum Sonenberg (McGill University). All of the experiments for the expression efficiency and the solubility were performed with C-terminally His_6-tagged proteins. Three mutation positions were selected as written in the results and discussion section. Mutations were introduced into the codon-optimized gene on pT7heIF2aopt with the Quick-change mutagenesis kit (Stratagene) one by one, which resulted in the plasmid pT7heIF2am3. Expression plasmid for the mutant was constructed in the same way as for the native protein.

Protein expression and solubility analysis

The BL21 Star (DE3) *E. coli* strain (Invitrogen) was used for all of the protein expression in this study. Transformed cells were cultured in LB or M9 media at 37 °C until the cell reached the mid-log phase, in which the OD at 600 nm of the culture was about 0.4. When ^2H_2O containing M9 was used, cells were cultured at 37 °C until the OD reached about 0.7. Subsequently, the target protein expression was induced with 0.5 mM IPTG at 20 °C for 6 h in LB, and for 9–20 hours in M9 media. When using chaperone co-expression, cells were doubly transformed with the expression plasmid for the target protein and the plasmid pG-Tf2, which contains the expression system for GroEL, GroES and TF (Nishihara et al., 2000). The plasmid pG-Tf2 was kindly donated by Dr Ohta. Expression of the chaperone proteins was induced with 10 ng ml^{-1} tetracycline at the beginning of the incubation in LB media. In the case of M9 media, the expression was induced with 5 ng ml^{-1} tetracycline 30 min before the target protein induction by IPTG.

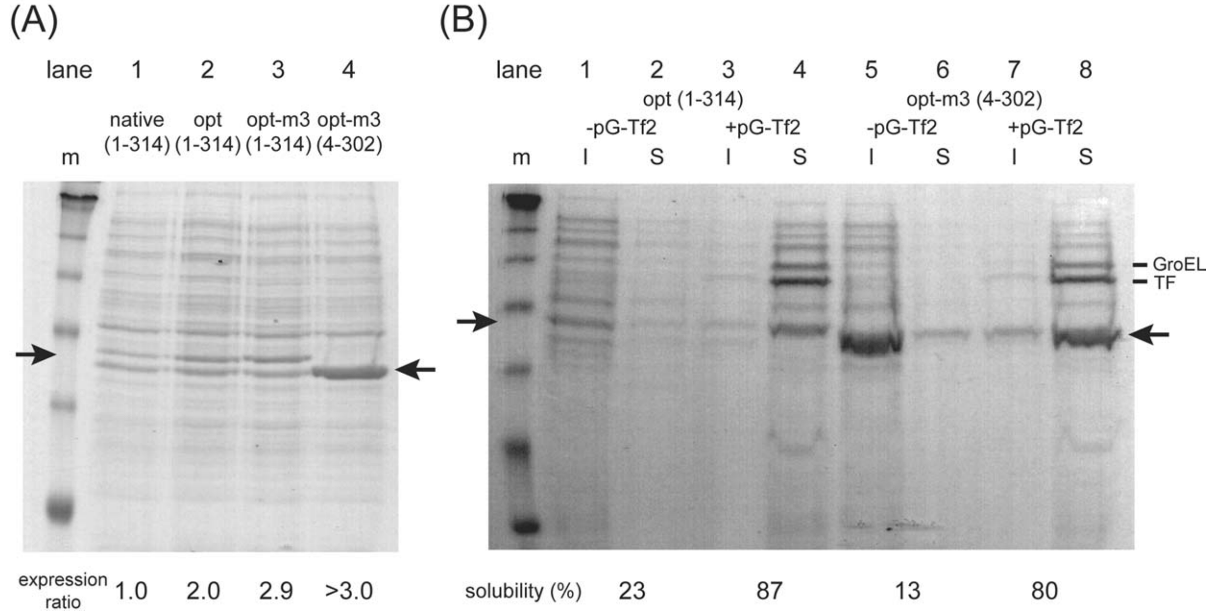

Figure 2. Effect of the codon-optimization and the co-expression of the chaperone proteins. (A) Overexpression was improved by codon-optimization and mutations. Total *E. coli* lysate with pET-heIF2anative(1-314) (lane 1), pET-heIF2aopt(1-314) (lane 2), pET-heIF2aopt-m3(1-314) (lane 3), and pET-heIF2aopt-m3(4-302) (lane 4) were analyzed by 12% SDS-PAGE. The band positions of the overexpressed heIF2α are indicated by arrows. The amount of the expressed heIF2α is quantified by the band intensities, and the ratios to that from pET-heIF2anative(1-314) are shown at the bottom. Since the band of overexpressed heIF2α from pET-heIF2aopt-m3(4-302) is clearly overlapped with the band of another protein, we just estimated the ratio at more than three. (B) Chaperone co-expression improved solubility. Insoluble (lanes 1, 3, 5 and 7) and soluble parts (lanes 2, 4, 6 and 8) of the total *E. coli* lysate, without (lanes 1, 2, 5, and 6) and with (lanes 3, 4, 7 and 8) co-expression of pG-Tf2 were analyzed by 12% SDS-PAGE for pET-heIF2aopt(1-314) (lanes 1–4) and pET-heIF2aopt-m3(4-302) (lanes 5–8). The positions of the bands for Gro EL and TF are indicated by bars and overexpressed heIF2α by arrows. Ratios of the soluble part to the total expressed heIF2α are shown at the bottom, analyzed by the band intensities.

Harvested cells were disrupted by sonication, and total cell lysates from originally equal volume cultures were analyzed by 12% SDS-PAGE for the expression analysis. For the solubility analysis, total cell lysates were separated by centrifugation into the soluble and the insoluble fractions. Originally equal volume samples of both fractions were analyzed by 12% SDS-PAGE. Protein bands on the gel were stained by Simply Blue SafeStain (Invitrogen) and band intensities were analyzed by Gel Doc system (Bio-Rad).

*Protein purification of heIF2*α

Cells were disrupted by sonication, and debris was eliminated by centrifugation. Supernatant was purified by Ni^{2+}-NTA agarose Super Flow according to the manufacturer's protocol (Novagen and Qiagen). Further purification was carried out by Gel filtration column chromatography with HiLoad Superdex 75 prep grade column (Pharmacia). The buffer for Gel filtration purification consists of 50 mM Na phosphate (pH 8.0), 350 mM Na_2SO_4, 2 mM DTT, and 1mM EDTA. The final yield of purified protein is typically about 40 and 15 mg from 1L culture of LB and M9 media, respectively.

NMR spectroscopy

The NMR sample was concentrated in the NMR buffer containing 50 mM Na phosphate (pH 7.0), 350 mM Na_2SO_4, 10 mM DTT, and 0.02% NaN_3. To obtain ^2H, ^{15}N, and/or ^{13}C- labeled proteins, ^2H_2O, $^{15}NH_4Cl$, and/or ^{13}C-labeled D-glucose were used as isotope sources, respectively. Amino acid-specific ^{15}N-labeled samples were prepared for Ala, Arg, Gly, Ile, Leu, Lys, Phe, Thr, Tyr and Val residues (LeMaster and Richards, 1985), using a BL21Star (DE3) strain. Two-dimensional ^1H-^{15}N HSQC spectra were recorded on a Bruker Avance 600 spectrometer, and all of the TROSY-based triple resonance spectra (Salzmann et al., 1998, 1999) and 3D ^{15}N-edited NOESY-HSQC spectrum were recorded on a Varian Inova 750 spectrometer. In order to obtain more sensitivity in TROSY-based measurements, C^{β} decoupling was employed during HNCA/HN(CO)CA (Matsuo et al., 1996a), a doubly selective R-SNOB pulse was used for CA/CO transfer (Matsuo et al., 1996b), and the overall duration of the pulse sequences was reduced as described by Loria et al. (1999). The data were processed with NMRPipe (Delaglio et al., 1995) and analyzed with NMRView (Johnson and Blevins, 1994). A list of chemical shifts has been deposited in the BioMagResBank (http://www.bmrb.wisc.edu) under accession number 5917.

Results and discussion

Synthesis of the codon-optimized artificial gene

Since the stable isotope labeled materials used for the preparation of the protein sample for NMR studies are expensive, it is desirable to construct an optimal expression system at the outset of a project. There are many reports which show that the difference of the codon usages makes it difficult to express large amounts of the eukaryotic proteins in prokaryotic cells, such as *E. coli* (Kane, 1995). Thus, we synthesized an artificial gene which consists of *E. coli* major codons for all amino acid residues. Based on the previous studies of codon usage, one major codon for *E. coli* is selected for one amino acid, GCT for alanine, CGT for arginine, and so on (Sharp and Li, 1987). Using these codons, we constructed the *E. coli* codon-optimized gene of heIF2α. At this stage, it is desirable to analyze possible RNA secondary structures of the corresponding mRNA in order to avoid unwanted RNA high-order structures which might affect translation by the *E. coli* system. If a highly stable RNA secondary structure is detected, selecting other codons could be desirable. This was not found to be necessary in our case.

Figure 1A shows a schematic representation of the synthetic method. Since heIF2α consists of 315 amino acid residues including an N-terminal methionine, the full DNA length of the gene is 948 bases, which contains a stop codon. We designed twelve 100 mer DNA template oligonucleotides to cover all DNA sequence as shown in Figure 1Ai. Complementary overlapping regions of 20–40 bases were designed at the 5′ and 3′ ends in the oligonucleotides as they anneal each other properly. In addition, two terminal primers were also designed for the final amplification as shown in Figure 1Aiii. The synthetic reaction was performed in two steps, an elongation step followed by an amplification step. The first step makes all oligonucleotides anneal on the right position and elongate their strands, which produced the template mixture for the second step. The second step is for amplification of the full length gene. This is similar to a PCR reaction, except that the template is not a single chain but consists of multiple chains. The products of these two steps were identified by agarose gel, as shown in Figure 1B. While the first elongation step produced a mixture of the various-length DNA (lane 1), the second amplification step produced one major DNA identified as a thick band in the gel (lane 2). This method is similar but not identical to the ‘assembly PCR’ described by Stemmer et al. (1995). Our procedure showed that the first PCR reaction in ‘assembly PCR’ can be replaced with a simple one-step elongation.

The amplified DNA was easily inserted into a cloning vector, pT7blue, using conventional methods. Although DNA sequencing identified some incorrectly polymerized bases especially at the bases next to the annealing regions, one of the four clones analyzed contained the correct target sequence, the *E. coli* codon-optimized gene of heIF2α. While two clones had a mutation at one point and one clone was mutated at two points, one had the correct sequence. We termed this plasmid pT7heIF2aopt.

Using pT7heIF2aopt, we constructed an expression plasmid for full-length heIF2α, pET-heIF2aopt(1-314). This was deriving from pET-30a(+) (Novagen) and results in an C-terminally His$_6$-tagged heIF2α. In order to confirm the efficiency of the codon-optimized gene, we also constructed the plasmid pET-heIF2anative(1-314) that expresses the same protein as pET-heIF2aopt(1-314) but contains the native cDNA sequence. A 1-L *E. coli* expression preparation appeared to produce enough material for several NMR samples when using pET-heIF2aopt(1-314), and heIF2α was the most abundant protein, as judged by 12% SDS-PAGE (lane 2 in Figure 2A). The pET-heIF2aopt(1-314) yielded at least two-fold more protein than pET-heIF2anative(1-314) (lane 1 in Figure 2A), as estimated by the band intensities.

Disappointingly, most of the overexpressed heIF2α was found in the insoluble fraction after the centrifugation following cell lysis (lane 1 in Figure 2B). Using different buffers for cell lysis did not increase the amount of soluble protein. We also tried several methods for refolding the protein from inclusion bodies but did not obtain sufficient amounts of soluble protein suitable for NMR studies.

Co-expression of chaperone proteins prevents the overexpressed heIF2α from forming inclusion bodies in the bacterial cell

Thus, in order to obtain more of the overexpressed heIF2α in the soluble fraction, we decided to co-express the protein with the prokaryotic chaperone proteins, GroEL-GroES and trigger factor (TF). GroEL and GroES are well-known prokaryotic molecular chaperones, which facilitate folding of proteins (Goloubinoff et al., 1989). TF is a chaperone-like factor that has peptidyl-prolyl *cis/trans* isomerase activity (Crooke and Wickner, 1987; Stoller et al., 1995). Nishihara et al. reported that co-expression of TF and GroEL-GroES was effective to improve the solubility of bacterially overexpressed recombinant proteins (Nishihara et al., 2000). Here, we used the previously described plasmid pG-Tf2 that overexpresses GroEL-GroES and TF under the control of the *Pxt-1* promoter inducible with tetracycline. The plasmid pG-Tf2 is a derivative of pACYC184, which is compatible with the ColE1 type plasmids used for most of the expression plasmids, and carries the chloramphenicol resistance gene as a selection marker. It is important to note that the strains in which the genes for minor tRNAs are supplied on the pACYC-type plasmid, such as *E. coli* Rosetta (Novagen) or *E. coli* CodonPlus (Stratagene), cannot be used with pG-Tf2, because the plasmid type and the selection marker are the same as pG-Tf2.

E. coli BL21 Star (DE3) was doubly transformed with pET-heIF2aopt(1-314) and pG-Tf2**,** and we compared the solubility of the overexpressed heIF2α obtained with and without pG-Tf2. As shown in Figure 2B, the solubility of heIF2α after the cell lysis dramatically improved with the co-expression of GroEL, GroES, and TF. The soluble portion of the totally expressed recombinant heIF2α was about 87% with the chaperones (lanes 3 and 4), while about 23% without them (lanes 1 and 2). We assume that this is due to the fact that the expressed chaperone proteins help proper folding of heIF2α, which is soluble, whereas incorrectly folded heIF2α forms aggregates in inclusion bodies.

Using these procedures we purified heIF2α after cell lysis from the soluble fraction using Ni-NTA affinity column and Superdex 75 Gel filtration column chromatographies. However, when attempting to concentrate purified heIF2α to levels needed for NMR assignments, most of the protein aggregated, and the highest concentration of soluble and stable protein we could reach was 0.05 mM. This was sufficient for recording 1D spectra and 2D ^{1}H-^{15}N HSQC spectra but not for assignments and structure determination and we concluded that wild-type heIF2α was not suitable for NMR structure determination.

Rational design of solubility-enhancing mutants

The plan to improve the solubility of our heIF2α construct was based on the hypothesis that there might be surface-exposed amino acids that cause low solubility but are not essential for correct folding and proper function of the protein. If we could identify these residues, it should be possible to replace them with amino acids that would enhance solubility. In order to identify candidates for mutation we relied on the homology of the N-terminal domain to a known structure of a domain from the ribosomal protein S1 (S1 domain), and on a sequence comparison of all known homologues of eIF2α. The ribosomal S1 protein contains the OB-fold, which consists of about 70 residues and contains five β-strands, forming an anti-parallel β-barrel (Arcus, 2002; Murzin, 1993). This fold has also been found in the RNA-binding domain of the *E. coli* polynucleotide phosphorylase (PPTase), which has recently been determined by NMR (Bycroft et al., 1997). A sequence alignment of S1 domains from PPTase, S1 protein, and heIF2α shows that these domains are sufficiently homologous to build a 3D model of the S1 domain of heIF2α (Figure 3A), and we used the PPTase structure to model the N-terminal S1 domain of heIF2α. To guide the design of mutations, we aligned sequences of eIF2α from several species using MAXHOM alignment of the PredictProtein server (http://cubic.bioc.columbia.edu/pp/). Although the sequences of the S1 domains of eIF2α are highly homologous, we identified three hydrophobic amino acid residues in human eIF2α (A27, L46 and V71) that are replaced with hydrophilic ones in the eIF2α homologues from *S. pombe* and/or *M. jannaschii* (Figure 3B). Inspection of the homology model for the N-terminal domain indicated that these hydrophobic amino acid residues are located on the protein surface, are hypothesized to be responsible for reduced solubility and might be exchanged with residues that would improve solubility without changing structure and function of the protein. This hypothesis is also supported by the fact that the corresponding amino acid residues in some S1 domains, indicated by the arrows in Figure 3A, are hydrophilic. Thus, we made a gene that contains the mutations A27Q, L46H, and

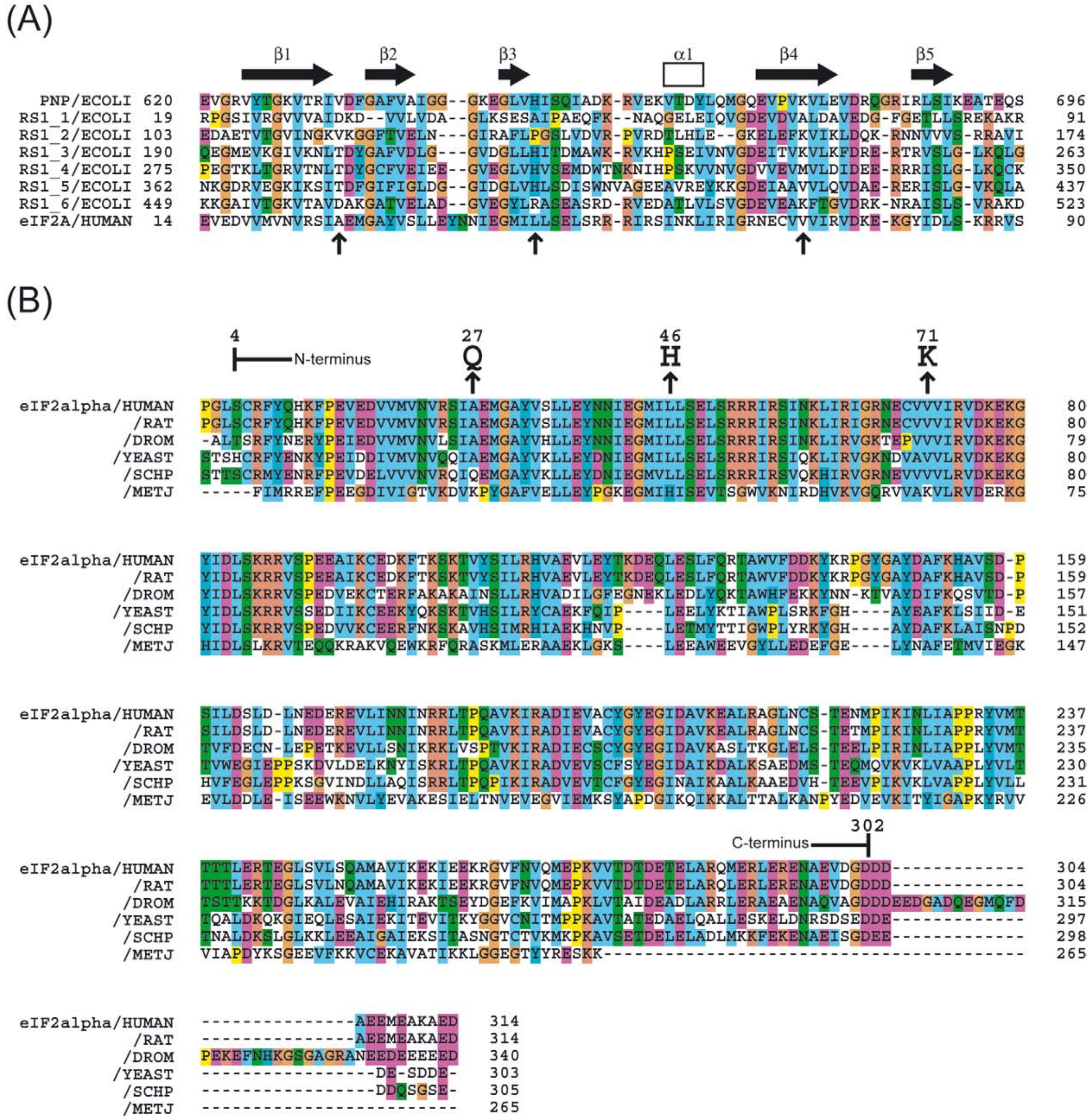

Figure 3. Primary sequence alignments of S1 domain and eIF2α. (A) Sequence alignment of S1 domain. Sequences of S1 domains of PPTase from *E. coli* (1st line), ribosomal S1 proteins from *E. coli* (2nd to 7th lines), and eIF2α from human (8th line) are shown. Secondary structure elements identified in PPTase (Bycroft et al., 1997) were shown at the top. The figure was created by the program ClustalX and colors are shown as a default of the program (Thompson et al., 1997). The mutation positions for the solubility-enhancement are indicated by arrows. (B) Primary sequence alignment of eIF2α. Sequences from human (*H. sapiens*), rat (*R. Norvegicus*), fly (*D. melanogaster*), baker's yeast (*S. cerevisiae*), fission yeast (*S. pombe*), and *M. jannaschii* were analyzed by ClustalX. Mutations introduced for the solubility-enhancement are shown at the top.

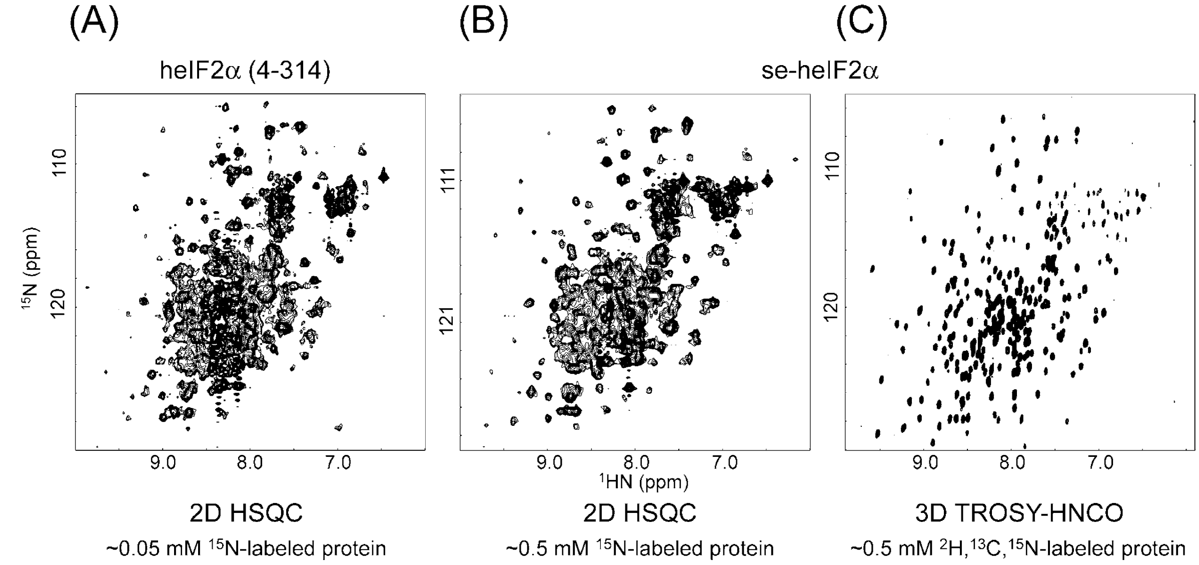

Figure 4. ^{1}H-^{15}N correlated spectra of heIF2α. (A) 2D ^{1}H-^{15}N HSQC spectrum of wild-type but N-terminally truncated ^{15}N-labeled heIF2α (4-314), at the maximum achievable concentration of ~0.05 mM. (B) 2D ^{1}H-^{15}N HSQC spectrum of ~0.5 mM ^{15}N-labeled se-heIF2α. (C) ^{1}H-^{15}N plane of the total projection of a 3D TROSY-HNCO spectrum of ~0.5 mM ^{2}H(~100%),^{13}C,^{15}N-labeled se-heIF2α.

V71K. A27Q is based on the sequence from *S. pombe*, and L46H and V71K are based on residues found in *M. jannaschii* (Figure 3B). In addition, we found three N-terminal residues (^{1}PGL^{3}) in heIF2α that are not conserved in some other species, such as *S. cerevisiae*, *S. pombe* or *M. jannaschii* and deleted them from our construct because of their hydrophobic character.

We also observed extremely strong peaks in the 2D ^{1}H-^{15}N HSQC spectrum for an earlier construct heIF2α (4-314), indicating that the C-terminus contained about 20 flexible residues (Figure 4A). This was also supported by a partial tryptic digestion experiment (data not shown). Considering all these findings, we finally constructed the expression vector, pET-heIF2aopt-m3(4-302), that produces the solubility-enhanced heIF2α (se-heIF2α). It contains mutations at the three positions described above, ranges from Ser4 to Asp302 (Figure 3B), and contains a C-terminal His$_6$-tag. We also made the expression vector pET-heIF2aopt-m3(1-314) in order to compare the effect on solubility of the mutants in the full-length construct.

Figure 2A shows that both mutations and truncation improved the amount of the expressed heIF2α. Mutations alone (lane 3) improved the expression about 1.5 times as much as non-mutated heIF2α (lane 2). Furthermore, additional N- and C-terminal truncation (lane 4) also improved the expression. To obtain overexpressed se-heIF2α in the soluble phase, chaperone co-expression is still necessary (Figure 2B, lane 5–8). While only 13% of the total expressed se-heIF2α was found in the soluble phase without co-expression of the chaperone proteins, 80% became soluble with co-expression.

As we had hoped, overexpressed se-heIF2α is highly soluble during purification and also sufficiently soluble to make an NMR sample of 0.5–0.7 mM concentration, in contrast to wild-type protein that is soluble only up to 0.05 mM. In addition, the sample is stable at least for a month (compared to a few days for the wild-type protein), enabling measurement of the necessary NMR spectra. The ^{1}H-^{15}N HSQC spectrum of se-heIF2α (Figure 4B) is similar to that of the native heIF2α (4-314) recorded at much lower concentration (Figure 4A). This fact indicates that the tertiary structure of se-heIF2α is conserved. Functional conservation is supported by the fact that se-heIF2α can be phosphorylated by the double-stranded RNA dependent protein kinase (PKR) (H. Aktas, pers. commun.).

The method presented here for designing solubility enhancing mutants is rather conservative and was based on earlier experience with improving the solubility of the adhesion domain of the glycoprotein CD58 (Sun et al., 1999). In that earlier study, six mutations had been introduced in an immunoglobulin superfamily fold. This made the mutant construct ex-

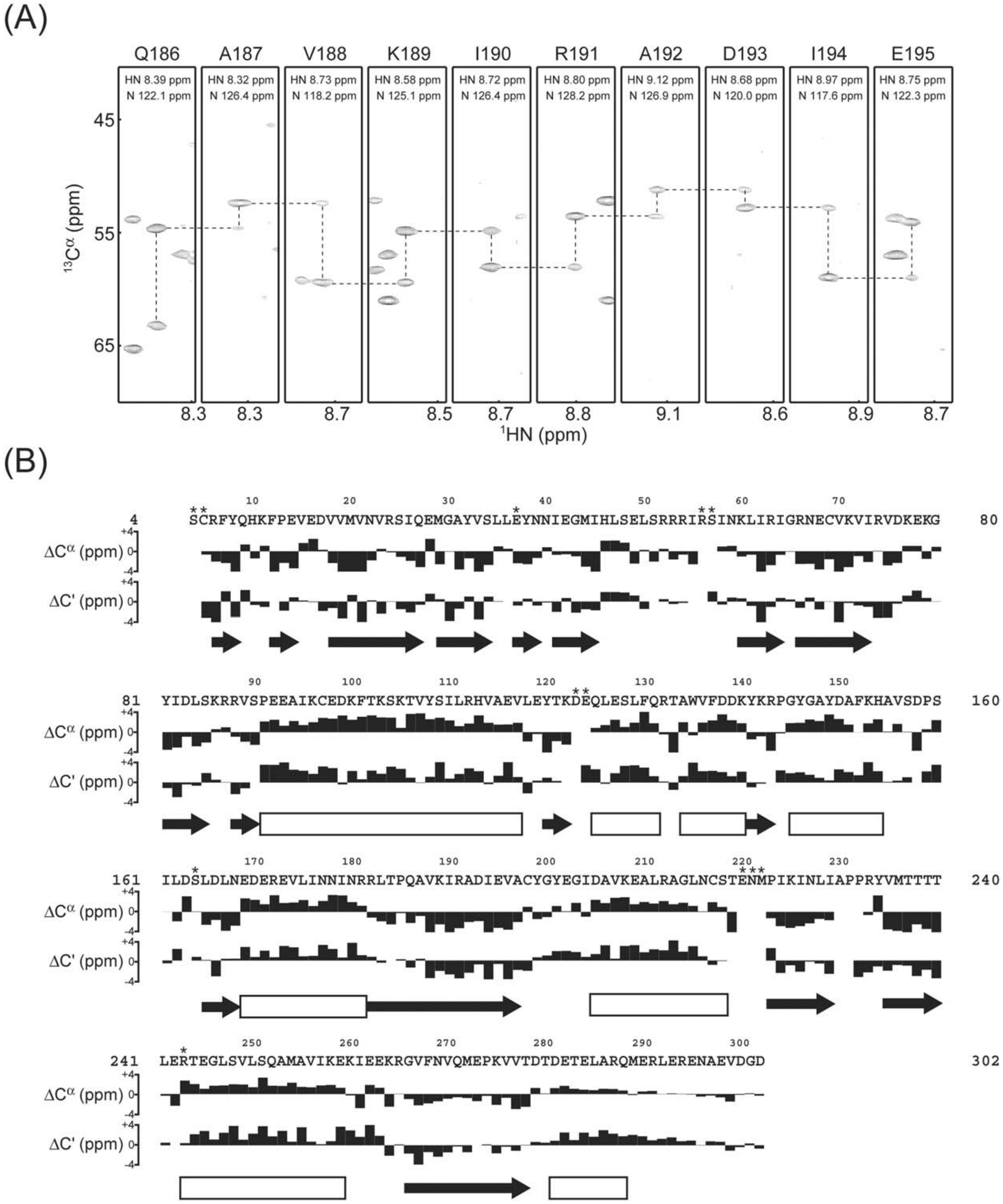

Figure 5. A summary of the backbone chemical shift assignment. (A) Strips form 3D TROSY-HNCA. A Strip plot from Q186 to E195 is shown and connectivity is indicated by dotted lines. (B) Deviations $^{13}C^{\alpha}$ and ^{13}CO from random coil chemical shifts. Unassigned HN groups are indicated by asterisk, and unassigned ^{13} C^{α} and ^{13}CO are by blank. Random coil chemical shift values are based on Wishart and Sykes (1994) and corrected for one-, two- and three-bond ^{2}H isotope effects (Venters et al., 1996). Secondary structure elements (arrows and boxes for strands and helices, respectively) identified by the CSI analysis (Wishart and Sykes, 1994) are shown at the bottom.

pressed in *E. coli* sufficiently soluble so that we could solve its structure by NMR. In contrast, the wild-type CD58 protein needs to be glycosylated for solubility, and is completely insoluble when expressed in *E. coli*. The mutant CD58 adhesion domain formed a soluble complex that was suitable for NMR and X-ray crystallography, and crystal structure of a complex with CD_2 was solved soon after (Wang et al., 1999). Mutations have been used in the past to enhance the solubility of proteins for crystallography studies, such as for the HIV 1 integrase (Jenkins et al., 1995). In that study the authors replaced systematically hydrophobic residues with hydrophilic ones and found that one of 29 mutations had a major and two others had minor solubility-enhancing effects. This differs from our approach where a single rationally planned set of mutations exhibited the desired effect.

The principle of solubility-enhancing mutation can also be employed in the absence of the knowledge of a tertiary fold. Comparison of homologous sequences often reveals segments of low conservation, which is suggestive of surface exposed loops. It seems possible that non-conserved hydrophobic amino acids located in such regions can be substituted with hydrophilic residues that are found in homologues to improve solubility without perturbing the structure of the target protein. This more aggressive approach could become a general avenue for solving the solubility problem of many poorly-behaved proteins.

Chemical shift assignment and secondary structure of se-heIF2α

The successful construction of the well-behaved se-heIF2α enabled us to pursue sequential resonance assignments. Figure 4C shows a total projection on the ^{1}H-^{15}N plane of a 3D TROSY-HNCO recorded with the 35 kDa ^{2}H, ^{13}C, ^{15}N-labeled se-heIF2α. We also measured TROSY-based 3D HN(CA)CO, 3D HNCA, 3D HN(CO)CA, 3D HN(CA)CB, and 3D HN(COCA)CB spectra. Figure 5A shows a strip plot of 3D TROSY-HNCA from Gln186 to Glu195. The quality of the 3D TROSY-type spectra was so high that we were able to assign more than 95% of the backbone (^{1}H^{N}, ^{15}N, ^{13}C$^{\alpha}$ and ^{13}CO) chemical shifts. Assignments were aided by analysis of 3D ^{15}N-edited NOESY-HSQC spectrum and 2D ^{1}H-^{15}N HSQC spectra of amino acid-specific ^{15}N-labeled samples. A summary of the assignment is shown in Figure 5B with deviations of ^{13}C$^{\alpha}$ and ^{13}CO chemical shifts from random coil shifts. CSI analysis (Wishart and Sykes, 1994) with ^{13}C$^{\alpha}$ chemical shifts revealed secondary structure elements of se-heIF2α as shown at the bottom line. NOE analysis indicated that the protein consists of a β strand-rich N-terminal OB domain (residues 4-90), an α-helix middle domain (residues 91-181), and an αβ-fold C-terminal domain (residues 182-302). These identified secondary structures are mostly consistent with the recently reported crystal structure of a fragment of heIF2α that contains the N-terminal and the middle domains (Nonato et al., 2002). The high quality of the spectra enabled us to determine the tertiary structure of se-heIF2α (manuscript in preparation). To achieve this we also needed to introduce ^{1}H-^{13}C labeled methyl groups of Ile, Leu and Val, together with protonated aromatic residues in a deuterated protein background (Gross et al., 2003; Medek et al., 2000).

Conclusion

We have used a rational strategy for improving expression level and solubility of the mammalian protein eIF2α that otherwise would not have been suitable for structural NMR studies. We synthesized an artificial gene to make an *E. coli* codon-optimized gene to increase expression levels. Co-expression of *E. coli* chaperone proteins from a separate plasmid increased the amount of target protein expressed in the soluble phase. In addition, incorporation of rationally designed solubility-enhancing mutations improved the solubility of the protein by an order of magnitude. With these methods, we obtained an *E. coli* overexpressed protein sample that was soluble up to 0.7 mM, which is sufficient for NMR structural studies. We have assigned more than 95% of the backbone chemical shifts of heIF2α and characterized its secondary structure. This approach may be generally applicable for NMR structural studies of biologically important but poorly behaved mammalian proteins.

Acknowledgements

We acknowledge Gregory Heffron and Dr John D. Gross for assistance of the measurements, and Dr Mallika Sastry for critically reading the manuscript. We thank Prof Nahum Sonenberg (McGill University) for providing the cDNA of heIF2α, and Dr. Tsunetaka Ohta (HSP Research Institute & Fujisaki Institute) for providing the chaperone-expressing plasmids.

This work was supported by the National Institute of Health (CA 68262, U19CA87427, GM47467 and RR-00995).

References

Arcus, V. (2002) *Curr. Opin. Struct. Biol.*, **12**, 794–801.

Bruggert, M., Rehm, T., Shanker, S., Georgescu, J. and Holak, T.A. (2003) *J. Biomol. NMR*, **25**, 335–348.

Bycroft, M., Hubbard, T.J., Proctor, M., Freund, S.M. and Murzin, A.G. (1997) *Cell*, **88**, 235–242.

Crooke, E. and Wickner, W. (1987) *Proc. Natl. Acad. Sci. USA*, **84**, 5216–5220.

Delaglio, F., Grzesiek, S., Vuister, G.W., Zhu, G., Pfeifer, J. and Bax, A. (1995) *J. Biomol. NMR*, **6**, 277–293.

Gardner, K.H. and Kay, L.E. (1997) *J. Am. Chem. Soc.*, **119**, 7599–7600.

Goloubinoff, P., Gatenby, A.A. and Lorimer, G.H. (1989) *Nature*, **337**, 44–47.

Goto, N.K., Gardner, K.H., Mueller, G.A., Willis, R.C. and Kay, L.E. (1999) *J. Biomol. NMR*, **13**, 369–374.

Gross, J.D., Gelev, V.M. and Wagner, G. (2003) *J. Biomol. NMR*, **25**, 235–242.

Hajduk, P.J., Augeri, D.J., Mack, J., Mendoza, R., Yang, J.G., Betz, S.F. and Fesik, S.W. (2000) *J. Am. Chem. Soc.*, **122**, 7898–7904.

Jenkins, T.M., Hickman, A.B., Dyda, F., Ghirlando, R., Davies, D.R. and Craigie, R. (1995) *Proc. Natl. Acad. Sci. USA*, **92**, 6057–6061.

Johnson, B.A. and Blevins, R.A. (1994) *J. Biomol. NMR*, **4**, 603–614.

Kane, J.F. (1995) *Curr. Opin. Biotechnol.*, **6**, 494–500.

LeMaster, D.M. and Richards, F.M. (1985) *Biochemistry*, **24**, 7263–7268.

Loria, J.P., Rance, M. and Palmer, 3rd, A.G. (1999) *J. Magn. Reson.*, **141**, 180–184.

Matsuo, H., Kupce, E., Li, H. and Wagner, G. (1996a) *J. Magn. Reson.*, **B113**, 91–96.

Matsuo, H., Kupce, E., Li, H. and Wagner, G. (1996b) *J. Magn. Reson.*, **B111**, 194–198.

Medek, A., Olejniczak, E.T., Meadows, R.P. and Fesik, S.W. (2000) *J. Biomol. NMR*, **18**, 229–238.

Morgan, W.D., Kragt, A. and Feeney, J. (2000) *J. Biomol. NMR*, **17**, 337–347.

Mueller, G.A., Choy, W.Y., Yang, D., Forman-Kay, J.D., Venters, R.A. and Kay, L.E. (2000) *J. Mol. Biol.*, **300**, 197–212.

Murzin, A.G. (1993) *EMBO J.*, **12**, 861–867.

Nishihara, K., Kanemori, M., Yanagi, H. and Yura, T. (2000) *Appl. Environ. Microbiol.*, **66**, 884–889.

Nonato, M.C., Widom, J. and Clardy, J. (2002) *J. Biol. Chem.*, **277**, 17057–17061.

Salzmann, M., Pervushin, K., Wider, G., Senn, H. and Wüthrich, K. (1998) *Proc. Natl. Acad. Sci. USA*, **95**, 13585–13590.

Salzmann, M., Wider, G., Pervushin, K., Senn, H. and Wüthrich, K. (1999) *J. Am. Chem. Soc.*, **121**, 844–848.

Sharp, P.M. and Li, W.H. (1987) *Nucl. Acids Res.*, **15**, 1281–1295.

Stemmer, W.P., Crameri, A., Ha, K.D., Brennan, T.M. and Heyneker, H.L. (1995) *Gene*, **164**, 49–53.

Stoller, G., Rucknagel, K.P., Nierhaus, K.H., Schmid, F.X., Fischer, G. and Rahfeld, J.U. (1995) *EMBO J.*, **14**, 4939–4948.

Sun, Z.Y., Dotsch, V., Kim, M., Li, J., Reinherz, E.L. and Wagner, G. (1999) *EMBO J.*, **18**, 2941–2949.

Thompson, J.D., Gibson, T.J., Plewniak, F., Jeanmougin, F. and Higgins, D.G. (1997) *Nucl. Acids Res.*, **25**, 4876–4882.

Venters, R.A., Farmer, 2nd, B.T., Fierke, C.A. and Spicer, L.D. (1996) *J. Mol. Biol.*, **264**, 1101–1116.

Wall, J.G. and Pluckthun, A. (1995) *Curr. Opin. Biotechnol.*, **6**, 507–516.

Wang, J.H., Smolyar, A., Tan, K., Liu, J.H., Kim, M., Sun, Z.Y., Wagner, G. and Reinherz, E.L. (1999) *Cell*, **97**, 791–803.

Wishart, D.S. and Sykes, B.D. (1994) *J. Biomol. NMR*, **4**, 171–180.