J Biomol NMR (2007) 37:53–63
DOI 10.1007/s10858-006-9103-0

ARTICLE

Fine-tuning of protein domain boundary by minimizing potential coiled coil regions

Naoko Iwaya · Natsuko Goda · Satoru Unzai · Kenichiro Fujiwara · Toshiki Tanaka · Kentaro Tomii · Hidehito Tochio · Masahiro Shirakawa · Hidekazu Hiroaki

Received: 8 August 2006 / Accepted: 5 October 2006 / Published online: 16 December 2006

Abstract Structural determination of individual protein domains isolated from multidomain proteins is a common approach in the post-genomic era. Novel and thus uncharacterized domains liberated from intact proteins often self-associate due to incorrectly defined domain boundaries. Self-association results in missing signals, poor signal dispersion and a low signal-to-noise ratio in ^{1}H–^{15}N HSQC spectra. We have found that a putative, non-canonical coiled coil region close to a domain boundary can cause transient hydrophobic self-association and monomer–dimer equilibrium in solution. Here we propose a rational method to predict putative coiled coil regions adjacent to the globular core domain using the program COILS. Except for the amino acid sequence, no preexisting knowledge concerning the domain is required. A small number of mutant proteins with a minimized coiled coil region have been rationally designed and tested. The engineered domains exhibit decreased self-association as assessed by ^{1}H–^{15}N HSQC spectra with improved peak dispersion and sharper cross peaks. Two successful examples of isolating novel N-terminal domains from AAA-ATPases are demonstrated. Our method is useful for the experimental determination of domain boundaries suited for structural genomics studies.

Keywords Domain boundary determination · Hydrophobic interaction · HSQC · Nonspecific self-association

Abbreviations

8-ANS	8-Anilino-1-naphthalenesulfonic acid
HSQC	Heteronuclear single quantum correlation spectroscopy
PCR	Polymerase chain reaction
AAA	ATPase associated with various cellular activities
KP60	katanin p60
NVL2	Nuclear VCP-like protein 2
VCP	Valosin containing protein p97
GST	Glutathione S-transferase

Introduction

An enormous amount of sequence information has been generated from numerous genome sequencing

Electronic Supplementary Material Supplementary material is available to authorised users in the online version of this article at http://www.dx.doi.org/10.1007/s10858-006-9103-0.

N. Iwaya · N. Goda · S. Unzai · K. Fujiwara · H. Tochio · M. Shirakawa · H. Hiroaki (✉)
Field of Supramolecular Biology, International Graduate School of Arts and Sciences, Yokohama City University, 1-7-29, Tsurumi, Yokohama, Kanagawa 230-0045, Japan
e-mail: hiroakih@tsurumi.yokohama-cu.ac.jp

N. Iwaya · H. Tochio · M. Shirakawa
Department of Molecular Engineering, Graduate School of Engineering, Kyoto University, Katsura, Kyoto 606-8501, Japan

T. Tanaka
Graduate School of Material Science, OMOHI-college, Nagoya Institute of Technology, Gokiso-cho, Nagoya 466-8555, Japan

K. Tomii
Computational Biology Research Center, The National Institute of Advanced Industrial Science and Technology, Aomi, Koto-ku, Tokyo 135-0064, Japan

projects (Wakeland and Wandstrat 2002) (also see http://www.ncbi.nlm.nih.gov/). The next major challenge is to focus on the genome-wide analysis of protein structure/function relationships (Burley and Bonanno 2003; Yokoyama et al. 2000; Phizicky et al. 2003). A sequence comparison of proteins from various genome sets has provided information concerning individual domains, which constitute an evolutionally conserved stretch of 50–300 amino acids capable of folding autonomously. Although many proteins are still annotated as ''function unknown'', an analysis of mammalian genomes shows that >70% of proteins are in fact ''multidomain proteins'', which harbor more than one protein domains (Vogel et al. 2004). Genetic dissection of genes encoding large proteins in order to obtain individual domains for structural studies is a key methodology in this field. Two major technical problems of working with individual domains in isolation from the parent protein have emerged: (i) the isolated domain tends to precipitate and is not well expressed in the cytosol of bacterial expression systems, and (ii) the isolated domain may form a soluble multimeric aggregate by nonspecific self-association. Both problems make the NMR analysis of novel protein domains difficult. When using solution NMR techniques, the first criterion of a promising protein domain for structure determination is to give a good HSQC spectrum. However, there is currently no rational approach to avoid self-association of uncharacterized protein domains.

In order to minimize self-association, we focused on regions of the target protein likely to form a coiled coil structure, which may act as a site for protein-protein interaction. Coiled coil structures are widely observed in proteins, and have been classified as one of the common supersecondary structures for protein-protein interactions (Yu 2002; Burkhard et al. 2001). Two to six amphipathic right-handed α-helices interact and wrap together to form a left-handed twist structure (Apic et al. 2001; Cohen and Parry 1990; Sanishvili et al. 2004). The periodic repeat of hydrophobic residues in an α-helical context is important for forming a coiled coil structure. Because seven amino acid residues form two rounds of each α-helix strand, the motif has a representative heptad repeat sequence with the amino acid residues designated *a* to *g* according to their position. Studies involving artificially designed coiled coils show that the *a*- and the *d*-positions are particularly important for inter-helical interactions and structural uniqueness. For example, the coiled coil without the structural uniqueness was found to be an ensemble of sub-optimal structures, resulting in a broadened NMR spectrum (Lumb and Kim 1995). The arrangement of the packing structure in the protein interior is an important *de novo* design target. According to these concepts, many successful designs for the control of the orientation (Oakley and Kim 1998) and the selective assembly of peptide fragments of coiled coils have been reported (Kiyokawa et al. 2000; Schnarr and Kennan 2003; Kashiwada et al. 2000). For example, the improved packing of the protein structure could lead to an increase in thermal stability (Sandberg and Terwilliger 1989). By contrast, substitution of one or more hydrophobic residues for small or charged residues can cause instability of a coiled coil structure as well as loss of structural uniqueness. This strategy can easily be used to eliminate predicted coiled coil regions of proteins, in which mutations are introduced to destabilize such packing.

Herein, we propose a simple prediction and design method to minimize protein self-association by destabilizing the putative coiled coil regions in the target protein in order to improve the HSQC signal with a minimum of experimental effort. Potential coiled coil regions are simply predicted by the program COILS (version 2.1) with the ''–mtidk'' option and weighting of hydrophobic residues (Lupas et al. 1991). The program takes the periodicity of hydrophobic amino acid residues in a heptad repeat into account for prediction. Using the ''–mtidk'' option, the program is more sensitive to the detection of non-canonical coiled coils. The putative coiled coil region is sometimes hidden in one domain region of the target protein. Since the program uses only the amino acid sequence as an input, we can virtually design any mutant sequence to avoid forming a coiled coil. Virtual mutants with deletion(s) and/or amino acid substitution(s) are first analyzed by COILS. A limited number of sequences for further study are then selected by using the COILS score as a guide. Using this approach, we have successfully isolated two N-terminal domains derived from the AAA-ATPases.

Materials and methods

Protein techniques

Vectors for the heterologous expression of GST fusion proteins of the N-terminal domains from both mouse and human katanin p60 (residues 1–90, denoted as $KP60_{1-90}$) and nuclear VCP-like protein 2 (residues 1–93, denoted as $NVL2_{1-93}$) were constructed using PRESAT-vector methodology, derived

from pGEX-4T3 vector (Amersham Biosciences, Piscataway, NJ) (Goda et al. 2004). Truncated constructs and the L73R mutant of mouse $KP60_{1-90}$ were prepared by site directed mutagenesis using GeneEditor (Promega, La Jolla, CA) according to the manufacturer's instructions. The solubility of GST-fusion proteins was assayed by the CDNB colorimetric assays according to the manufacturer's instructions (Amersham Biosciences). The ^{15}N-labeled recombinant proteins for NMR spectroscopy were generated in *E. coli* BL21(DE3) from a 1.0 l M9 minimal medium culture grown in the presence of $^{15}NH_4Cl$ as the sole nitrogen source at 30°C. The cell lysate after sonication was cleared by centrifugation and then applied to a DEAE-Sepharose column (Amersham Biosciences), and then affinity purified by Glutathione Sepharose (Amersham Biosciences) chromatography. The GST tag was removed by thrombin ''on-beads'', and the protease was trapped using benzamidine Sepharose (Amersham Biosciences) and then dialyzed.

NMR Spectroscopy

Samples for NMR spectroscopy contained mouse KP60 N-terminal domains at a concentration of approximately 0.1 mM in 5% D_2O–95% H_2O, 20 mM sodium phosphate buffer (pH 7.5) and 1% CHAPS. Samples for NMR spectroscopy contained mouse NVL2 N-terminal domains at a concentration of approximately 0.1 mM in 5% D_2O–95% H_2O and 25 mM sodium phosphate buffer (pH 6.4). ^1H–^{15}N HSQC spectra for KP60 N-terminal domains were recorded on a 500 MHz Bruker DRX NMR spectrometer equipped with a cryogenic probe at 25°C. ^1H–^{15}N HSQC spectra for NVL2 domains were recorded on a 800 MHz Bruker *Avance* NMR spectrometer equipped with a cryogenic probe at 25°C. Data were processed by using NMRPipe (Delaglio et al. 1995).

CD and fluorescence spectroscopy

CD spectra of mouse KP60 N-terminal domains were measured in 0.1-cm path length cuvettes at 25°C using a JASCO J-720W spectropolarimeter (JASCO, Co, Tokyo). 10 μM of each protein was dissolved in buffer containing 1 mM EDTA and 50 mM Tris–HCl (pH 7.5). Fluorescence spectra of 8-ANS bound to mouse KP60 N-terminal domains and bovine α-lactalbumin were measured in a 1-cm path length cuvette at 25°C using a Shimadzu RF-5300PC spectrofluorophotometer (Kyoto, Japan). An excitation wavelength of 370 nm was used and emission from 400 to 600 nm measured. Fluorescence enhancement experiments were done by measuring the difference spectra of fluorescence emission from various concentrations of 8-ANS (0–100 μM) with and without 4 μM of protein in buffer containing 1 mM EDTA and 50 mM Tris–HCl (pH 7.5), except α-lactalbumin (pH 2.0). Because the unit of fluorescence intensity is arbitrary, α-lactalbumin at pH 2.0 was used as a reference of fluorescence enhancement of a molten globule protein.

Analytical ultracentrifugation

Sedimentation velocity experiments were carried out using an Optima XL-I analytical ultracentrifuge (Beckman Coulter, Fullerton, CA) with a Beckman An-50 Ti rotor. For sedimentation velocity experiments, cells with a standard Epon two-channel centerpiece and sapphire windows were used. Sample (400 μl) and reference buffer (420 μl) were loaded into cells. The rotor temperature was equilibrated at 20°C in the vacuum chamber for 1–2 h prior to start-up. Absorbance (OD_{280}) scans were collected at 10 min intervals during sedimentation at 50×10^3 rpm. The sedimentation velocity experiments for KP60 N-terminal domains were conducted at concentrations of between 0.17 and 0.4 mg/ml. Partial specific volume of the protein, solvent density and solvent viscosity were calculated from standard tables using the program SEDNTERP, version 1.08 (Laue et al. 1992). The resulting scans were analyzed using the continuous distribution ($c(s)$) analysis module in the program SEDFIT version 9.3 (Schuck et al. 2002). Sedimentation coefficient increments of 50 or 100 were used in the appropriate range for each sample, and the frictional coefficient was allowed to float during fitting. The weight average sedimentation coefficient was obtained by integrating the range of sedimentation coefficients in which peaks were present.

Sedimentation equilibrium experiments were also carried out in cells with six channel centerpiece and quartz windows. The sample concentrations were 0.17, 0.29 and 0.4 mg/ml. The absorbance wavelength was set at 280 nm, and data was acquired at 20°C. Data were obtained at 15, 20, 25, and 30×10^3 rpm. A total equilibration time of 14 h was used for each speed, with a scan taken at 12 h to ensure equilibrium had been reached. Data analysis was performed by global analysis of data sets obtained at different loading concentrations and rotor speeds using XL-A/XL-I Data Analysis Software Version 4.0.

Results

Initial NMR assessment of N-terminal domains from AAA ATPases

We have analyzed members of the AAA-ATPase family of proteins (Beyer 1997; Lupas et al. 2002); katanin p60 (KP60) and nuclear VCP-like protein 2 (NVL2). We anticipated that the N-terminal region of AAA-ATPases may possess modular domains responsible for specific substrate and/or adaptor binding regions. Our recent success in determining the structure of the N-terminal domain(s) from PEX1 ATPase (Shiozawa et al. 2004) encouraged us to apply such an approach to the AAA proteins. Figure 1 shows a sequence alignment of the N-terminal 100 amino acid region of KP60 and NVL2. Starting from the full length sequence of mammalian proteins of both KP60 and NVL2, orthologous sequences were retrieved from the nr (non-redundant) protein sequence database using PSI-BLAST. Although there had been no indication of the presence of an isolatable domain at the N-terminus of either protein at the starting point of this research, the initial domain boundaries were defined as residues 1–90 and 1–93 for mouse KP60 and mouse NVL2, respectively, because these positions are less conserved among their orthologs (Fig. 1). Note that the newer version of fold recognition servers (e.g. FORTE and FUGUE) have been enabled to detect a MIT-domain like region at the N-terminus of KP60, after the structures of MIT domain were reported (Ciccarelli et al. 2003; Scott et al. 2005; Takasu et al. 2005). GST-fusion expression vectors for both domains of mouse and human orthologs were constructed, and the domains were expressed in *Escherichia coli* BL21(DE3). Preliminary experiments showed that mouse $KP60_{1–90}$ and mouse $NVL2_{1–93}$ were more soluble than the human orthologs (data not shown). Thus, the recombinant mouse proteins were selected for further investigation.

The $^{1}H–^{15}N$ HSQC spectrum of mouse $KP60_{1–90}$ is shown in Fig. 2e. We anticipated a total of 95 main chain amide signals for $KP60_{1–90}$ (including six additional vector-derived signals), but only about 60 signals were obtained. The dispersion of observed signals was similar to that of a typical HSQC pattern for a folded molecule, suggesting certain parts of the domain were folded. We performed an extensive search of different buffer conditions for NMR measurements of $KP60_{1–90}$ according to standard practice (reviewed in Kremer and Kalbitzer 2001), by varying pH (5.5–7.5) and ionic strength (0–0.5 M), prior to applying the method described below (Fig. S1 of Supplementary Material). Because the quality of HSQC spectra was largely independent of salt conditions, self-association appeared to be hydrophobic in nature. Nevertheless, more than 30 cross peaks were still missing. Thus, we attempted to improve the spectral quality of these protein domains using an alternative approach such as introducing mutations.

Fine-tuning of boundaries of mouse KP60 N-terminal domain by destabilizing putative coiled coil region

The sequences of mouse and human $KP60_{1–90}$ were analyzed by the program COILS with window sizes of 14, 21 and 28 residues. These are the default values of

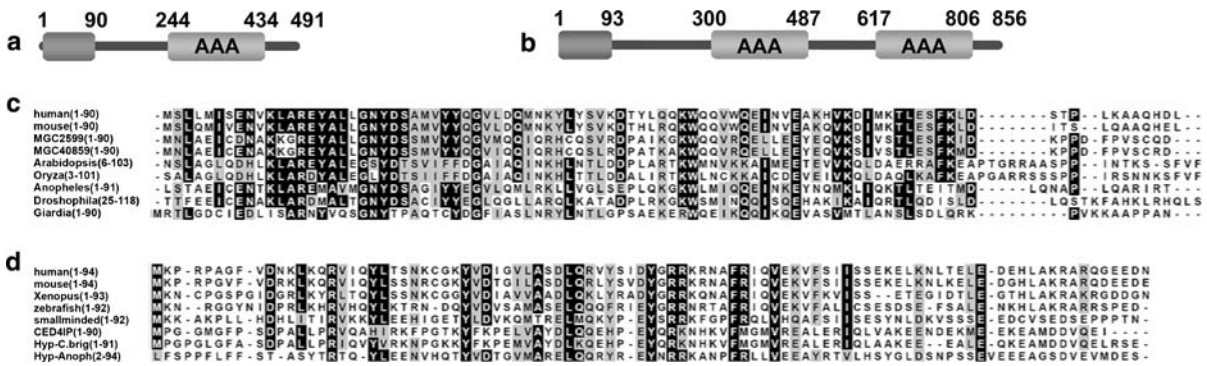

Fig. 1 Multiple alignment of protein sequences used in this study. (**a**) Domain architectures of mouse katanin p60 (KP60). (**b**) Domain architectures of mouse nuclear VCP-like protein 2 (NVL2). (**c**) Multiple alignment of the N-terminal region of KP60 orthologs; human (UniProtKB accession no. O75449) and mouse (Q9WV86) katanins, human (MGC2599; Q9BW62) and mouse (MGC40859; Q8K0T4) hypothetical proteins, *Arabidopsis* (Q9SEX2), *Oryza* (Q8S118), *Anopheles* (Q7PY77, fragment), *Drosophila* (Q9VN89) and *Giardia* (Q7R5W7) katanin p60. (**d**) Multiple alignment of the N-terminal region of NVL2 orthologs; human (O15381), mouse (Q9DBY8), *Xenopus* (Q7ZXI4), zebrafish (Q803I9), *Drosophila* (smallminded; P91638), *C. elegans* CED4IP (Q9U8K0), *C. briggsae* (CBG20797; Q60SQ4) and *Anopheles* (ENSANGP00000024422; Q7PID7) hypothetical proteins. Residue numbers are in parentheses. The sequence alignments were generated by ClustalX

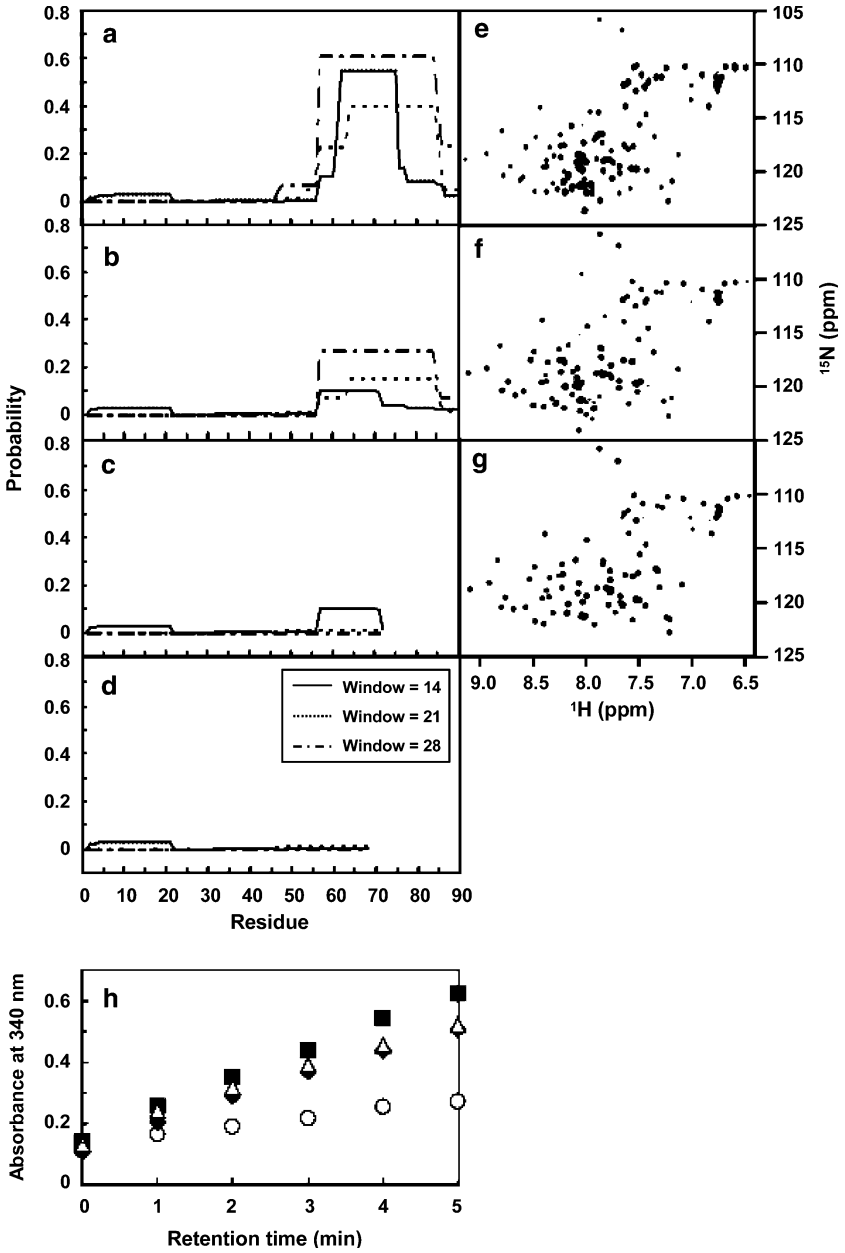

Fig. 2 Correlation of coiled coil propensity and ^{1}H–^{15}N HSQC spectral dispersion for the N-terminal regions of katanin p60. (**a**–**d**) Coiled coil propensity calculated by the program COILS. **a**: $KP60_{1–90}$, **b**: $KP60_{1–90}$ (L73R), **c**: $KP60_{1–72}$, **d**: $KP60_{1–68}$, solid, dotted and dashed lines indicate window sizes of 14, 21 and 28 residues, respectively. (**e**–**g**) ^{1}H–^{15}N HSQC spectra. **e**: $KP60_{1–90}$, **f**: $KP60_{1–90}$ (L73R), **g**: $KP60_{1–72}$. (**h**) Solubility of GST-tagged katanin p60 N-terminal regions monitored by CDNB colorimetric assays. Filled diamond: $KP60_{1–90}$, filled box: $KP60_{1–90}$ (L73R), open triangle: $KP60_{1–72}$, open circle: $KP60_{1–68}$

the size of gliding window to calculate the coiled-coil propensity (Lupas et al. 1991). In all cases, approximately 30 amino acid residues at the C-terminal region of the putative domain showed a coiled coil propensity of greater than 50% (Fig. 2a). Thus far, however, the predicted coiled coil region has not been ascribed any biological function. The length of the predicted coiled coil region was consistent with the number of missing cross peaks in the HSQC spectra (Fig. 2e). We therefore assumed that the missing signals on the HSQC spectrum originated from the region of the predicted coiled coil.

We designed several virtual mutants for the putative N-terminal domain of KP60. The virtual sequences for the N-terminal domain of KP60 were as follows: $KP60_{1–90}$ (I69R), $KP60_{1–90}$ (I80R), $KP60_{1–90}$ (L73R), $KP60_{1–90}$ (L78R), $KP60_{1–90}$ (V66R), $KP60_{1–75}$, $KP60_{1–72}$ and $KP60_{1–68}$. It is known that increasing the number of heptad repeats dramatically stabilizes the coiled coil in two-, three-and four-stranded coiled coils (Harbury et al. 1993), and vice versa. It is known that partial digestion by endogenous trypsin-like protease in *E. coli* often generates products with Lys or Arg as the C-terminal residue. Thus, we considered several truncated mutants with Lys or Arg at the C-terminus, which may avoid the heterogeneity of the C-terminal residues. The substitution mutants involved replacing a hydrophobic residue with an arginine, which was expected to

increase solubility. In general, hydrophobic residues, such as Ile, Val, Leu, Met, Phe, Tyr and Trp, can be substituted by either charged (Arg, Lys, Asp, Glu) or small (Ser, Gly) residues. All the sequences were subjected to COILS analysis, and those that gave a substantial reduction of coiled coil propensity ($KP60_{1–90}$ (L73R), $KP60_{1–72}$, $KP60_{1–68}$) were chosen for further study (Fig. 2b–d).

The three candidate proteins were engineered for expression and purified. The shortest construct $KP60_{1–68}$ tended to precipitate during protein expression and purification. The solubility of the KP60 N-terminal domains was semi-quantitatively assessed by CDNB colorimetric assays of GST activity (Fig. 2h). Interestingly, the $KP60_{1–90}$ (L73R) mutant exhibited greater solubility than intact $KP60_{1–90}$. The $KP60_{1–90}$ (L73R) and $KP60_{1–72}$ mutants were soluble, and were further analyzed by $^{1}H–^{15}N$ HSQC (Fig. 2f and g, respectively; expansions of these spectra are available in Fig. S2). More than 60 NH signals were detected, which were essentially identical among all three constructs. However, the intensities in $KP60_{1–72}$ were more uniform than those of $KP60_{1–90}$ and $KP60_{1–90}$ (L73R). As a result, $KP60_{1–72}$ was chosen for further NMR structural determination studies. By further optimization of the solvent conditions for $KP60_{1–72}$, 98% completeness of the HSQC signal data was achieved (Fig. S3). Using this data, we were able to identify the missing signals in the spectra of $KP60_{1–90}$. The signals originating from 73 to 90 and additional seven signals (Y22, E62, A63, Q65, V66, K67 and I69) were absent, and six out of the seven missing signals were originating from the helical region adjacent to the residues 73–90. These residues are thought to be involved in the self-association interface, which is consistent with our initial assumption.

The putative coiled coil region adopts a α-helix, but is not sensitive to 8-ANS

On the basis of the good signal dispersion of the HSQC spectra, it is likely that $KP60_{1–72}$ is the core folded domain. Thus, we assessed whether the truncated region, residues 73–90, adopts a α-helical conformation as predicted. The CD spectra of $KP60_{1–90}$ and $KP60_{1–90}$ (L73R) showed a substantial increase of α-helicity compared to that of $KP60_{1–72}$ (Fig. 3), suggesting the presence of a helix within the putative coiled coil region. The mutant $KP60_{1–90}$ (L73R) exhibited reduced self-association whilst retaining the helical structure in the residues 73–90. In contrast, the spectra of $KP60_{1–68}$ showed a drastic loss of helical content, suggesting that the deletion of only four terminal residues introduced disturbance of the proper folding of the core domain. Interestingly, the GST-fusion form of $KP60_{1–68}$ was less soluble, which may reflect decreased stability.

It is known that a protein in a molten-globule state often shows broadening or elimination of signals in $^{1}H–^{15}N$ HSQC spectra, probably because of chemical exchange between conformationally heterogeneous sub-optimal species (Schulman et al. 1997). Since the signals originating from residues 73–90 were not fully observed, $KP60_{1–90}$ may fit this criterion. Proteins in a molten-globule state are known to be sensitive to 8-ANS and increase their fluorescence. Thus, we examined whether the observed weak self-association of $KP60_{1–90}$ through the putative coiled coil region is observable in the 8-ANS fluorescence enhancement assay (Fig. 4). We measured 8-ANS fluorescence of various concentrations (0–100 μM) in the presence of 4 μM KP60 N-terminal domains. The same experiment was performed with α-lactalbumin instead of KP60 N-terminal domains as a positive control for a molten globule protein. Nevertheless, as shown in Fig. 4, only a small fluorescence enhancement of $KP60_{1–90}$ as well as $KP60_{1–90}$ (L73R) was observed. This shows that the completeness and the signal dispersion of $^{1}H–^{15}N$ HSQC spectra of the protein of interest is a better diagnostic fingerprint for the degree of transient self-association than hydrophobic fluorescent probes.

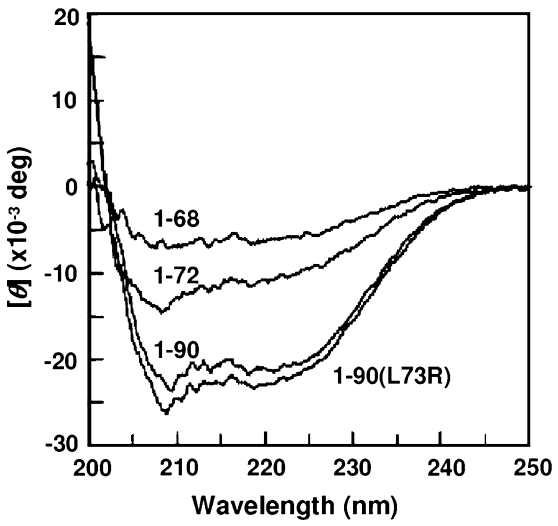

Fig. 3 Comparison of CD spectra of KP60 N-terminal domains. N-terminal domain variants are shown in the panel. Each 10 μM of protein were dissolved in buffer containing 1 mM EDTA and 50 mM Tris–HCl (pH 7.5)

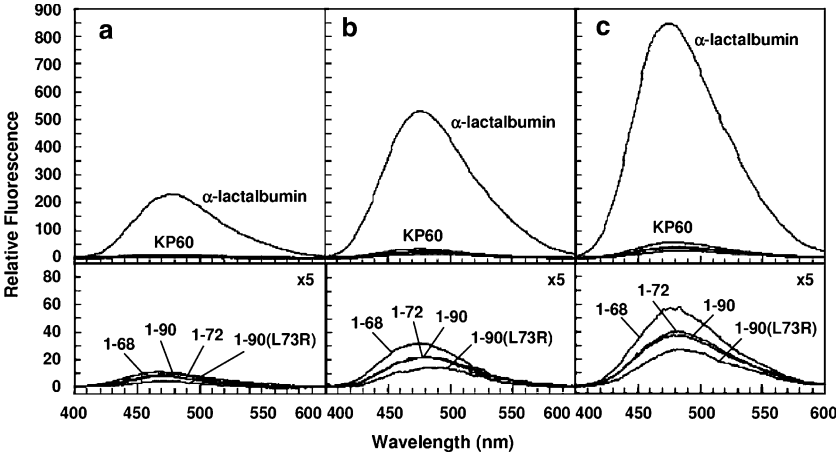

Fig. 4 Fluorescence enhancement experiments of 8-ANS bound to KP60 N-terminal domains and α-lactalbumin. The difference spectra of fluorescence emission between ANS with and without the protein are shown. (**a**) 10 μM, (**b**) 20 μM, (**c**) 40 μM of ANS were added to 4 μM of protein in buffer containing 1 mM EDTA and 50 mM Tris–HCl (pH 7.5), except α-lactalbumin (pH 2.0). Lower panel is × 5 magnification of upper panel. KP60 N-terminal domain variants are shown in the spectra

The putative coiled coil region caused monomer–dimer equilibrium in solution

Sedimentation velocity provides hydrodynamic information about the sample and establishes the size distribution of proteins due to their different rates of migration in the centrifugal field. We analyzed the sedimentation velocity of $KP60_{1–90}$, $KP60_{1–90}$ (L73R) and $KP60_{1–72}$ at three different protein concentrations (0.17, 0.29 and 0.4 mg/ml) in order to assess whether self-association was also concentration-dependent. We analyzed continuous distribution $c(s)$ versus sedimentation coefficient for each data set by using the program SEDFIT, because $c(s)$ distributions provide excellent sensitivity and resolution, enabling a clear distinction between different sedimenting species.

Figure 5 shows the distributions of sedimentation coefficients on KP60 N-terminal domains at the concentration of 0.4 mg/ml. First, we confirmed that $KP60_{1–72}$ was a monomer. The $c(s)$ distribution of $KP60_{1–72}$ shows the presence of a single species in the solution with a sedimentation coefficient (s) of 1.2 (± 0.1)S. The $c(s)$ distribution of $KP60_{1–72}$ did not show any significant change upon varied protein concentration. The molecular mass of $KP60_{1–72}$ was determined to be 9.7 (± 0.1) kDa by the sedimentation equilibrium experiments (Fig. S4), which agreed very well with the theoretical value (9.35 kDa) based on the amino acid sequence.

However, the $c(s)$ distribution of $KP60_{1–90}$ and $KP60_{1–90}$ (L73R) showed that they were not simple monomeric proteins. $KP60_{1–90}$ gave a single peak, but the shape was much broader and the peak position much larger compared to that of $KP60_{1–72}$ (Fig. 5). At increased protein concentration (0.17, 0.29 and 0.4 mg/ml), the weight average of the peak increased (1.7, 1.8 and 1.9 S, data not shown). These results are consistent with a relatively rapid reversible equilibrium between the monomer and oligomer species, which was previously assumed from NMR spectra. To further examine the oligomerization status of $KP60_{1–90}$, we performed sedimentation equilibrium experiments (Fig. S4). Sedimentation equilibrium is a good indication of a thermodynamic equilibrium of the self-association systems and can be used to determine the dissociation constants (reviewed in Lebowitz et al. 2002; Laue and Stafford 1999). We applied several models for fitting. A simple monomeric model gave an estimated molecular mass of 20.1(± 0.2) kDa, which was greater than the theoretical value of 11.4 kDa. In the case of $KP60_{1–90}$ (L73R), $c(s)$ distribution profiles also showed protein concentration-dependent peaks at positions between that of $KP60_{1–72}$ and $KP60_{1–90}$. The profile has a main peak relatively close to the peak for $KP60_{1–72}$, with a larger S-value component as a shoulder (Fig. 5). Upon increasing protein concentration the shoulder became larger. Thus, we conclude that $KP60_{1–90}$ (L73R) exists as a monomer–dimer (or oligomer) equilibrium, although the level of oligomerization was much less than observed for $KP60_{1–90}$.

Application of the COILS method for fine-tuning of boundaries of mouse NVL2 N-terminal domain

The successful determination of the domain boundary for KP60 encouraged us to apply the same approach to

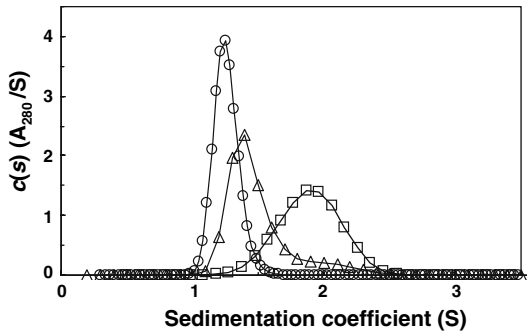

Fig. 5 Distribution of sedimentation coefficients [$c(s)$] for KP60 N-terminal domains. Calculated $c(s)$ is plotted versus sedimentation coefficient (s). Open box: $KP60_{1–90}$, open triangle: $KP60_{1–90}$ (L73R), open circle: $KP60_{1–72}$. Experiments were conducted at an initial protein concentration of 0.4 mg/ml in 1 mM EDTA, 100 mM NaCl and 20 mM sodium phosphate buffer (pH 7.5, 20°C) and a rotor speed of 50×10^3 rpm, and data was collected at time intervals of 10 min. The calculated values for the weight-average sedimentation coefficient (s) are s = 1.9 S, 1.4 S and 1.2 S for $KP60_{1–90}$, $KP60_{1–90}$ (L73R) and $KP60_{1–72}$, respectively

another case, NVL2. $^1H–^{15}N$ HSQC spectra of mouse $NVL2_{1–93}$ are shown in Fig. 6e. As with KP60, only about 50 signals were obtained from an expected total of 94 amide signals for $NVL2_{1–93}$ (including four additional vector-derived signals). The sequence of $NVL2_{1–93}$ was analyzed by the program COILS (Fig. 6a). Improved constructs for the N-terminal domain of NVL2 were designed using a similar approach to that described for KP60. The sequences $NVL2_{1–82}$, $NVL2_{1–77}$ and $NVL2_{1–74}$ were analyzed by COILS (Fig. 6b–d). In this case we only focused on deletion of the predicted coiled coil region, and the three constructs were engineered for expression and purified. HSQC data for all these mutant proteins showed a dramatic improvement in terms of either the number of observed peaks or the line widths compared to that of $NVL2_{1–93}$ (Fig. 6e–h; expansions of these spectra are available in Fig. S5). Interestingly, several weaker and sharper signals for $NVL2_{1–82}$ and $NVL2_{1–77}$, within the range of ^1H chemical shift of 7.6-8.4, were not observed in the spectra for $NVL2_{1–93}$ and $NVL2_{1–74}$. These signals were presumably from the C-terminal region, which exists as a random coil, rather than as a result of self-association. Therefore it would appear that removal of residues 83–93 is sufficient to hinder self-association.

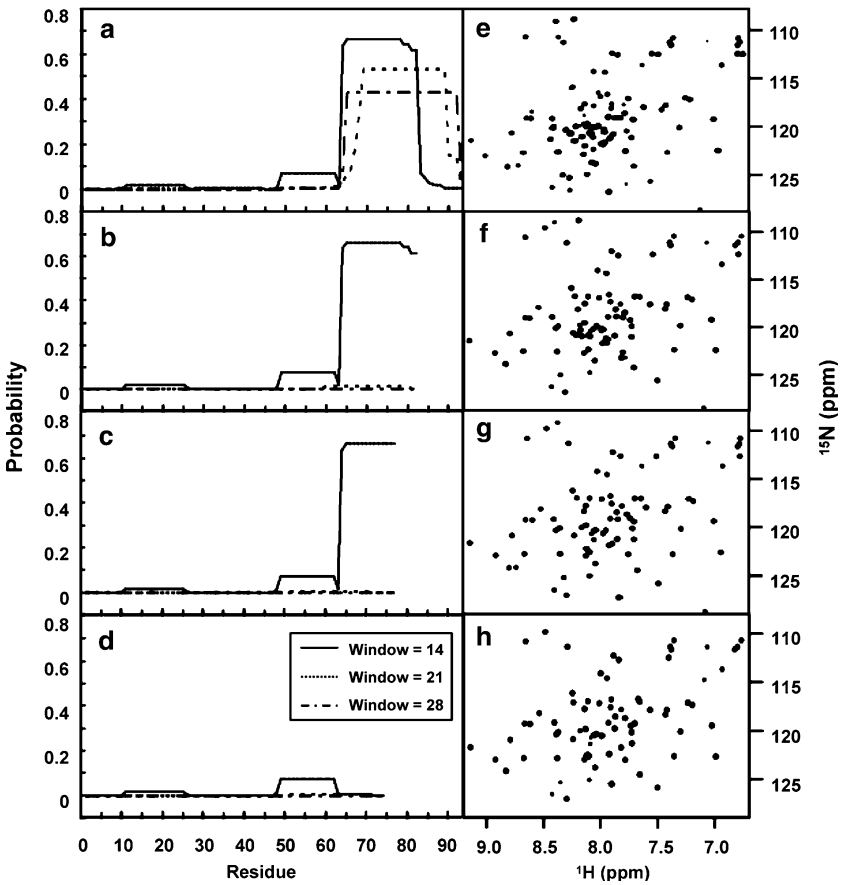

Fig. 6 Correlation of coiled coil propensity and $^1H–^{15}N$ HSQC spectral dispersion for the N-terminal regions of nuclear VCP-like protein 2. (**a**–**d**) Coiled coil propensity calculated by the program COILS. **a**: $NVL2_{1–93}$, **b**: $NVL2_{1–82}$, **c**: $NVL2_{1–77}$, **d**: $NVL2_{1–74}$, solid, dotted and dashed lines indicate window sizes of 14, 21 and 28 residues, respectively. (**e**–**h**) $^1H–^{15}N$ HSQC spectra. **e**: $NVL2_{1–93}$, **f**: $NVL2_{1–82}$, **g**: $NVL2_{1–77}$, **h**: $NVL2_{1–74}$

Discussion

We have shown two successful examples of rational fine-tuning of protein domain boundaries, KP60 and NVL2, excised from multidomain proteins, in which the initial HSQC assessment of the domains was poor. Our methodology circumvents the laborious practice of improving the protein characteristics, which usually involves engineering and screening large numbers of mutants in order to obtain a promising HSQC signal. This process is simplified by identifying residues likely to cause self-association prior to mutant design. We have employed a primitive bioinformatics approach to help eliminate the problem of self-association.

The putative N-terminal domains of KP60 and NVL2 gave ^{1}H–^{15}N HSQC spectra typical of proteins displaying a tendency to self-associate. The disappearance of approximately 30 amide proton signals originating from interfacial residues upon oligomerization was attributed to exchange broadening at equilibrium. Interestingly, although the signals arising from the interfacial residues disappeared, there was no substantial line-broadening of other signals, suggesting that the coiled coil segments are connected to the core of the folded domain via a flexible segment. The proteins may have been in equilibrium between a monomer and oligomer with an uncertain association number. If a monomer–dimer self-association equilibrium model is applied, the dissociation constant of the equilibrium can be obtained from the analytical ultracentrifugation data. The molecular mass was constrained to the theoretical value and the dissociation constant was allowed to float during fitting. The fits were good (Fig. S4) and the dissociation constants were calculated to be (10^{-5}~10^{-6}) M and (10^{-3}~10^{-4}) M for $KP60_{1–90}$ and $KP60_{1–90}$ (L73R), respectively. This means that when the protein concentration is set to 0.1 mM as in an NMR sample, the monomer–dimer ratio of the samples was approximately 1:2~7 and 1:0.1~0.5, respectively. Thus, the dimer species is dominant for $KP60_{1–90}$ in the NMR conditions, whereas the monomer is dominant for $KP60_{1–90}$ (L73R). In both cases, chemical exchange between monomer and dimer spoiled the HSQC spectra. The result obtained from both NMR and analytical ultracentrifugation was consistent with the proposed monomer–dimer equilibrium.

Although analytical ultracentrifugation studies showed the molecular weight average of $KP60_{1–90}$ was close to that of a dimer, we do not rule out the potential formation of oligomers. Amino acid residues occurring in the (*a*–/*d*–) positions of the predicted coiled coil regions are different from the ideal amino acids for a homotypic coiled coil formation. Specifically, either (Leu/Leu) or (Ile/Leu) in the (*a*–/*d*–) positions respectively, form a dimeric coiled coil and the combination of (Ile/Ile) or (Leu/Ile) affords trimer or tetramer formation, respectively (Harbury et al. 1993). Since neither the sequences of KP60 and NVL2 met these criteria, the oligomerization through the putative coiled coil regions is probably promiscuous.

Comparison of the HSQC spectra of the putative coiled coil-containing domains with those of the pruned domains show that the peak positions of dispersed amide proton signals are essentially identical. This supported the conclusion that the putative coiled coil regions are only engaged in inter-molecular interactions rather than internal contacts. The putative coiled coil regions of $KP60_{1–90}$ (L73R) adopt a α-helical conformation (Fig. 3), with decreased self-association (Fig. 5). The truncated constructs appear to encode the minimal core structural domains, which are monomeric in nature and therefore useful for further structure determination.

Based on the two successful examples of this study, we propose a flow diagram for obtaining a protein domain suitable for structural NMR studies (Fig. 7). Firstly, the boundaries of putative globular domains are identified by several bioinformatics techniques. The highly conserved region among the orthologs is tentatively defined as a domain. In addition, sequences likely to fall outside the putative domain are predicted as either disordered regions (e.g., disopred2) (Ward et al. 2004) or ''linker''-like regions (e.g., DomCut) (Suyama and Ohara 2003). Alternatively, if the target protein exhibits high or substantial similarity to known

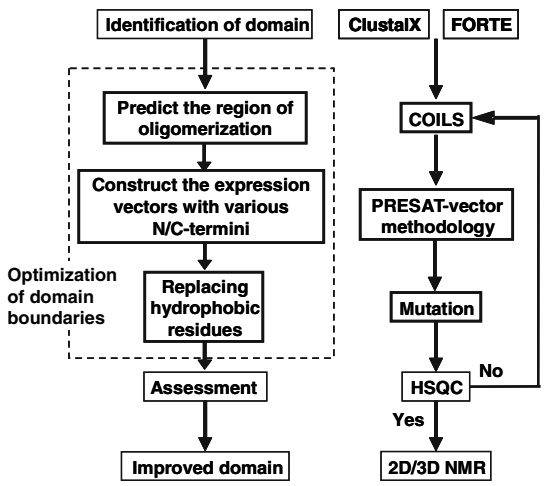

Fig. 7 Flow-chart for optimizing domain boundaries by minimizing coiled coil propensity. The steps drawn by bold lines are skipped if an initial HSQC assessment is attempted first

protein domains, the initial domain boundary is defined according to the known example by using the PSI-BLAST (Altschul et al. 1997), Pfam (Bateman et al. 2002), SMART (Schultz et al. 1998) or ProDom (Bru et al. 2005) servers. In addition, fold recognition algorithms, such as FORTE (Tomii and Akiyama 2004) and FUGUE (Shi et al. 2001), are equally helpful because they are highly sensitive in finding distant homologs. Secondly, the sequence is subjected to the program COILS with the ''-mtidk'' option. In most cases, a putative coiled coil region is found at the N-or C-termini of the prototype domain. Protein sequences with either deletion or amino acid substitution at the coiled coil region are then virtually generated. The virtually generated sequences are analyzed by COILS, and coiled coil propensity is calculated. It is better to monitor all the scores from window sizes of 14, 21, and 28 residues, because the sensitivities may differ. Thirdly, the expression plasmids with the designed domain boundaries are constructed, by using high-throughput PRESAT-vector methodology (Goda et al. 2004). A limited number of candidate target domains are cloned in parallel and incorporated into a bacterial expression vector, such as a GST-fusion vector. Amino acid substitution may also be introduced. Finally, only the most soluble proteins are subjected to ^{15}N-labeling studies to obtain ^{1}H–^{15}N HSQC spectra. In this study, we performed the initial HSQC assessment for both the prototypical domains of KP60 and NVL2 according to standard practice. However, we propose the COILS analysis could be performed prior to the assessment by HSQC spectra (Fig. 7).

Ideally, a protein sample for 3D-structure determination should be stable, highly soluble and should not undergo self-association. The development of high-throughput structural genomics requires the adoption of novel strategies to obtain suitable protein samples. Examples include parallel construction of different fusion proteins (Hammarstrom et al. 2002), cell-free expression systems using PCR-amplified DNA fragments (Sawasaki et al. 2002), and parallel *E. coli* protein expression in a 96-well plate format (Shih et al. 2002). These candidate proteins were isotopically-labeled and subjected to HSQC measurements. Finally, a ''go or no-go'' decision of the selected protein as a target for structure determination is made based on the quality of the HSQC spectra. Constructs giving the best quality HSQC spectral data are ranked for further analysis. Such a strategy is suited to maximize throughput for genome-wide structural studies (Christendat et al. 2000). In contrast, our strategy is a rational method that proposes to improve protein behavior in solution using ''negative design'' of potential coiled coils. Additionally, we have launched an automatic web-based server for this design method (available at http://www.mbs.cbrc.jp/coiled-coil/). Adoption of our strategy might avoid discarding biologically important protein samples with less promising HSQC spectra at an early stage. Furthermore, this design approach could be useful for preparing protein constructs for X-ray crystallography, because the putative coiled coil regions may prevent crystallization. In conclusion, it is useful to examine the existence of putative coiled coil regions associated with the target protein domain, when the initial HSQC spectrum is poor, or even before measuring the initial HSQC. Furthermore, our method is widely applicable because it does not require any preexisting knowledge of a ''not-yet-characterized'' domain.

Special note

The domain boundary optimizing server is available at http://www.mbs.cbrc.jp/coiled-coil. Access is freely available after publication of this manuscript.

Acknowledgments This work was partly supported by grants to H.H. from the Japanese Ministry of Education, Science, Sports and Culture (Protein3000), and was supported by grants to H.H. and K.T. from Japan Science and Technology Agency (BIRD). We thank Mr. K. Inomata for help with data representation.

References

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Nucleic Acids Res 25:3389–3402
Apic G, Gough J, Teichmann SA (2001) J Mol Biol 310:311–325
Bateman A, Birney E, Cerruti L, Durbin R, Etwiller L, Eddy SR, Griffiths-Jones S, Howe KL, Marshall M, Sonnhammer EL (2002) Nucleic Acids Res 30:276–280
Beyer A (1997) Protein Sci 6:2043–2058
Bru C, Courcelle E, Carrere S, Beausse Y, Dalmar S, Kahn D (2005) Nucleic Acids Res 33 Database Issue: D212–D215
Burkhard P, Stetefeld J, Strelkov SV (2001) Trends Cell Biol 11: 82–88
Burley SK, Bonanno JB (2003) Methods Biochem Anal 44:591–612
Christendat D, Yee A, Dharamsi A, Kluger Y, Gerstein M, Arrowsmith CH, Edwards AM (2000) Prog Biophys Mol Biol 73:339–345
Ciccarelli FD, Proukakis C, Patel H, Cross H, Azam S, Patton MA, Bork P, Crosby AH (2003) Genomics 81:437–441
Cohen C, Parry DA (1990) Proteins 7:1–15
Delaglio F, Grzesiek S, Vuister GW, Zhu G, Pfeifer J, Bax A (1995) J Biomol NMR 6:277–293
Goda N, Tenno T, Takasu H, Hiroaki H, Shirakawa M (2004) Protein Sci 13: 652–658
Hammarstrom M, Hellgren N, van Den Berg S, Berglund H, Hard T (2002) Protein Sci 11:313–321

Harbury PB, Zhang T, Kim PS, Alber T (1993) Science 262:1401–1407
Kashiwada A, Hiroaki H, Kohda D, Nango M, Tanaka T (2000) J Am Chem Soc 122:212–215
Kiyokawa T, Kanaori K, Tajima K, Tanaka T (2000) Biopolymers 55:407–414
Kremer W, Kalbitzer HR (2001) Methods Enzymol 339:3–19
Laue TM, Shah BD, Ridgeway TM, Pelletier SL (1992) In: Harding S, Rowe A, Horton J (eds) Analytical ultracentrifugation in biochemistry and polymer science. Royal Society of Chemistry, Cambridge, UK, pp 19–125
Laue TM, Stafford III WF (1999) Annu Rev Biophys Biomol Struct 28:75–100
Lebowitz J, Lewis MS, Schuck P (2002) Protein Sci 11:2067–2079
Lumb KJ, Kim PS (1995) Biochemistry 34:8642–8648
Lupas A, Van Dyke M, Stock J (1991) Science 252:1162–1164
Lupas AN, Martin J (2002) Curr Opin Struct Biol 12:746–753
Oakley MG, Kim PS (1998) Biochemistry 37:12603–12610
Phizicky E, Bastiaens PI, Zhu H, Snyder M, Fields S (2003) Nature 422:208–215
Sandberg WS, Terwilliger TC (1989) Science 245:54–57
Sanishvili R, Pennycooke M, Gu J, Xu X, Joachimiak A, Edwards AM, Christendat D (2004) J Struct Funct Genom 5:231–240
Sawasaki T, Ogasawara T, Morishita R, Endo Y (2002) Proc Natl Acad Sci USA 99:14652–14657
Schnarr NA, Kennan AJ (2003) J Am Chem Soc 125:667–671
Schuck P, Perugini MA, Gonzales NR, Howlett GJ, Schubert D (2002) Biophys J 82:1096–1111
Schulman BA, Kim PS, Dobson CM, Redfield C (1997) Nat Struct Biol 4:630–634
Schultz J, Milpetz F, Bork P, Ponting CP (1998) ProcNatl Acad Sci USA 95:5857–5864
Scott A, Gaspar J, Stuchell-Brereton MD, Alam SL, Skalicky JJ, Sundquist WI (2005) Proc Natl Acad Sci USA 102:13813–13818
Shi J, Blundell TL, Mizuguchi K (2001) J Mol Biol 310:243–257
Shih YP, Kung WM, Chen JC, Yeh CH, Wang AH, Wang TF (2002) Protein Sci 11:1714–1719
Shiozawa K, Maita N, Tomii K, Seto A, Goda N, Akiyama Y, Shimizu T, Shirakawa M, Hiroaki H (2004) J Biol Chem 279:50060–50068
Suyama M, Ohara O (2003) Bioinformatics 19:673–674
Takasu H, Jee JG, Ohno A, Goda N, Fujiwara K, Tochio H, Shirakawa M, Hiroaki H (2005) Biochem Biophys Res Commun 334:460–465
Tomii K, Akiyama Y (2004) Bioinformatics 20:594–595
Vogel C, Berzuini C, Bashton M, Gough J, Teichmann SA (2004) J Mol Biol 336:809–823
Wakeland EK, Wandstrat AE (2002) Curr Opin Immunol 14:622–626
Ward JJ, McGuffin LJ, Bryson K, Buxton BF, Jones DT (2004) Bioinformatics 20:2138–2139
Yokoyama S, Hirota H, Kigawa T, Yabuki T, Shirouzu M, Terada T, Ito Y, Matsuo Y, Kuroda Y, Nishimura Y, Kyogoku Y, Miki K, Masui R, Kuramitsu S (2000) Nat Struct Biol 7 Suppl:943–945
Yu YB (2002) Adv Drug Deliv Rev 54:1113–1129