J Struct Funct Genomics
DOI 10.1007/s10969-011-9116-0

A simplified recipe for assigning amide NMR signals using combinatorial ^{14}N amino acid inverse-labeling

Hidekazu Hiroaki · Yoshitaka Umetsu · Yo-ichi Nabeshima · Minako Hoshi · Daisuke Kohda

Received: 27 January 2011 / Accepted: 2 August 2011

Abstract Assignment of backbone amide proton resonances is one of the most time-consuming stages of any protein NMR study when the protein samples behave non-ideally. A robust and convenient NMR procedure for analyzing spectra of marginal-to-low quality is helpful for high-throughput structure determination. The ^{14}N selective- and inverse-labeling method is a candidate solution. Here, we present a simplified protocol for assigning protein backbone amide NMR signals. When ^{14}N inversely labeled residues are present in a protein, their backbone NH cross peaks vanish from the protein's ^{1}H–^{15}N HSQC spectrum, and thus, their chemical shifts can be readily identified by a process of elimination. Some metabolically related amino acids, for example, Ile, Leu, and Val, cannot be individually incorporated but can be inversely labeled together. We optimized and simplified the protocol and M9-based medium formula for the ^{14}N selective- and inverse-labeling method without any additives. Our approach should be cost-effective, because the method could be additively applied stepwise, even when the proteins of interest were found to be non-ideal.

Keywords Combinatorial inverse-labeling · Aβ(1–40) peptide · NMR sample preparation · Isotope labeling

Abbreviations

HSQC	Heteronuclear single quantum coherence spectroscopy
IPTG	Isopropyl-β-D-galactoside
IL1β	Interleukin-1β
SOFAST-HMQC	Band-selective optimized flip-angle short-transient heteronuclear multiple quantum coherence spectroscopy
BEST	Band-selective excitation short-transient
TROSY	Transverse relaxation optimized correlation spectroscopy

Introduction

The development of a robust and cost-effective NMR method that is applicable to assignment of backbone amide proton resonances of proteins with non-ideal properties is a challenge in structural genomics research. Usually, assignment of the backbone resonances is one of the most time-consuming but indispensable stages of any NMR study. A ^{1}H–^{15}N HSQC spectrum (and its variants, for

Electronic supplementary material The online version of this article (doi:10.1007/s10969-011-9116-0) contains supplementary material, which is available to authorized users.

H. Hiroaki (✉) · Y. Umetsu
Division of Structural Biology, Kobe University Graduate School of Medicine, 7-5-1, Kusunoki-cho, Chuo, Kobe 650-0017, Japan
e-mail: hiroakih@med.kobe-u.ac.jp

H. Hiroaki · Y. Umetsu
The Structural Biology Resreach Center and Division of Biological Science, Graduate School of Science, Nagoya University, Furo-cho, Chikusa-ku, Nagoya 464-8601, Japan

H. Hiroaki · D. Kohda
Department of Structural Biology, Biomolecular Engineering Research Institute, 6-2-3 Furue-dai, Suita, Osaka 565-0874, Japan

Y. Nabeshima · M. Hoshi
Department of Pathology and Tumor Biology, Kyoto University Graduate School of Medicine, Kyoto 606-8501, Japan

D. Kohda
Medical Institute for Bioregulation, Kyushu University, 3-1-1 Higashi-ku, Hakata, Fukuoka 812-8582, Japan

Published online: 25 August 2011

example, SOFAST-HMQC [1] and TROSY [2]) provides a fingerprint of the protein of interest and is thereby important as a reference spectrum for other triple resonance experiments. The information obtained from a ^{1}H–^{15}N HSQC spectrum is not only necessary for 3D structure determination but also indispensable for examining protein dynamics, H/D exchange, and protein–ligand interactions. A process for assignment of the backbone resonances of proteins with molecular weights as large as 20 kDa routinely starts with preparation of a uniformly $^{13}C/^{15}N$-labeled non-deuterated protein. Using this sample, several pairs of 3D NMR experiments are recorded, for example, HNCA/HN (CO)CA, HNCACB/CBCA (CO)NH, and HNCO/HN (CA)CO [3–5]. Among these pairs, one provides information on the intraresidual spin systems while another provides details of the interresidual connectivities. Thus, analysis of these data sets leads to sequence-specific assignment of all main chain signals along the peptide sequence.

Nevertheless, non-ideal properties of a protein sample often hamper the analysis of a protein's NMR data set because some key resonances are often missing. Such undesired behaviors include limited solubility, nonspecific self-association, chemical exchange, and internal motion of the protein. These properties often result in low signal-to-noise ratios of certain cross-peaks in 3D NMR experiments, for example, those arising from $^{13}C\beta$s in HNCACB and CBCA (CO)NH. As a result, assignment of the backbone resonances becomes difficult or even impossible because connectivities between residues are not uniquely identified. In addition, some non-ideal experimental conditions provide similar results. Examples of non-ideal NMR conditions of physiological or biophysical interest include highly viscous conditions, NMR analysis of a protein inside a living cell, presence of lipid micelles, and a high level of denaturants or salts.

At present, there are several options for solving the problem of missing resonances: (a) use a SOFAST-/BEST-NMR pulse scheme to gain signal intensity per experimental time [1, 6]; (b) use a nonlinear, sparse sampling method for the indirect dimension of a 3D/4D NMR experiment to minimize the measurement time and to maximize S/N [7–9]; and (c) use protein samples that contain selectively ^{15}N-labeled residues [10–12]. However, there exist many drawbacks associated with these solutions in terms of the costs. For example, using the third strategy, the cost and effort associated with sample preparation is a major concern when projects like structural genomics that require massive protein preparation and NMR data acquisition are involved. The cost and effort required to introduce selectively labeled amino acids by using cell-free translation systems are now relatively minimized, but are still relevant [12].

In this study, we describe a simple method to obtain the information necessary for residue-type assignments that uses "combinatorial inverse-labeling" of sets of specific amino acids, an idea that was initially proposed by Shortle [13]. In the original method, MOPS minimal media [14], modified such that glycerol was the sole carbon source, was employed with the combinatorial inverse-labeling method. A matrix of four combinations of amino acids redundantly covering 17 amino acids was designed and inversely labeled. The resulting five HSQC data sets (four inversely labeled and one uniformly labeled) were processed with the "labeling pattern matrix" to identify each amino acid type independent from the other 3D experiments. One major drawback of this method is the requirement of a special processing application.

Here, we further optimized the method for high-throughput structural determination in the current manner. By this method, we prepared ^{15}N uniformly labeled and ^{14}N selectively and inversely labeled protein samples as recombinant proteins expressed in *Escherichia coli* BL21(*DE3*) that were grown in M9 minimal media containing $^{15}NH_4Cl$ as the sole nitrogen source with or without supplementation of unlabeled (^{14}N) amino acids. In addition, we proposed some new sets of combinations of metabolically related amino acids, which were found to be useful.

Materials and methods

Protein techniques

The *E. coli* expression plasmid for IL1β, which has an added N-terminal methionine and whose synthesis is under the control of the p*tac* promoter, was generously provided by Dr. Y. Kikumoto [15]. ^{15}N-labeled recombinant IL1β and ^{14}N selectively and inversely labeled IL1βs were expressed in *E. coli* BL21(*DE3*) that had been transformed with the plasmid. Cells were grown at 30°C in 0.1 L M9 minimal media containing $^{15}NH_4Cl$ as the sole nitrogen source with or without unlabeled amino acid(s). Protein production was induced by adding IPTG to a final concentration of 1 mM when the OD600 of a culture reached 0.4. Cells were harvested after 4 h. The details of this expression protocol are outlined in Table 1. Each cell-free extract was applied to a S-Sepharose fast flow column (GE-Healthcare Biosciences), and the IL1β was eluted using a NaCl gradient containing 20 mM sodium phosphate (pH 6.5).

The *E. coli* expression plasmid for human Aβ(1–40) peptide, which has a N-terminal maltose-binding protein tag and a linker containing tobacco etch virus protease cleavage site followed by Aβ(1–40), was generously

Table 1 The procedure for ^{14}N selective- and inverse-labeling of a protein

1. Prepare ^{15}N M9 medium containing glucose (4 g/L), ^{15}N-ammonium chloride (0.7 g/L), NaH_2PO_4–$12H_2O$ (15 g/L), K_2HPO_4 (3 g/L), NaCl (0.5 g/L), $MgSO_4$ (0.24 g/L), $CaCl_2$-H_2O (15 mg/L), vitamins, nucleobases (each at 100 mg/L), and ampicillin (50 mg/L)
2. Pick three to five colonies of *E. coli* BL21(*DE3*) to add into 1 mL LB medium and grow the culture overnight at 30–37°C
3. Add 1 mL of the bacterial culture to 50 mL M9 medium containing $^{15}NH_4Cl$ and incubate at 30–37°C until the OD600 of the culture is 0.8–1.0 (pre-culture medium)
4. Add the pre-culture medium to 450 mL of M9 medium containing $^{15}NH_4Cl$ and incubate at 30–37°C until the OD600 of the culture is 0.2–0.4 (main culture medium)
5. Add 100 mg/L (final concentration) of each 14N selected-labeled amino acid* and incubate the culture at 30–37°C for an additional 30–60 min
6. Induce protein expression with the addition of 0.5–1.0 mM IPTG (final concentration)
7. Incubate the culture at 30–37°C for an additional 2–4 h
8. Harvest the cells

* Use of a stock solution of 20 mg/mL (~pH 10) is recommended for Tyr and Phe. Other amino acids can be added in powder form

provided by Dr. D. Hamada. ^{15}N-labeled recombinant Aβ(1–40) and ^{14}N selectively and inversely labeled Aβ(1–40)s were expressed in *E. coli* BL21(*DE3*) that had been transformed with the plasmid. Cells were grown at 37°C in 0.1 L M9 minimal media containing $^{15}NH_4Cl$ as the sole nitrogen source with or without unlabeled amino acid(s). Protein production was induced by adding IPTG to the final concentration of 1 mM when the OD600 of a culture reached 0.45. ^{14}N-amino acids were added 45 min prior to IPTG induction. Cells were harvested after 18 h at 20°C in order to prevent aggregation of Aβ(1–40). Each cell-free extract was applied to an amylose resin column (New England Biolabs), and the fusion protein was eluted by maltose. The fusion protein was treated with tobacco etch virus protease (New England Biolabs), and the resultant peptide was finally purified by reversed-phase HPLC using a linear acetonitrile gradient containing dilute ammonium hydroxide.

NMR spectroscopy

Samples used for NMR spectroscopy were approximately 0.5 mM and 0.1 mM in fully $^{13}C/^{15}N$-labeled and inversely labeled protein, respectively, 5% D_2O–95% H_2O, and 20 mM sodium phosphate (pH 5.5). For assignment of the backbone resonances, initially, HNCA, HN (CO)CA, HNCACB, CBCA (CO)NH, HNCO, HN (CA)CO, and 3D ^{15}N-edited NOESY-HSQC spectra of fully $^{13}C/^{15}N$-labeled proteins were recorded at 25°C using a 600 MHz Bruker DMX600 spectrometer or a 600 MHz Bruker *AVANCE*-III spectrometer [1]. Then, ^1H–^{15}N HSQC spectra or ^1H–^{15}N SOFAST-HMQC spectra of several inversely labeled proteins were recorded for residue-type assignment according to the backbone resonance assignment circumstances. Data were processed using NMRPipe [16]. IL1β backbone signal assignments were taken from the literature [17]. Aβ(1–40) backbone signal assignments were taken from the literature [18] and confirmed by this method.

Results and discussion

To obtain the information necessary for backbone signal assignments when using data sets derived from low-quality 3D NMR spectra, we developed an assignment strategy using a "combinatorial inverse-labeling" method. According to this method, we prepared ^{15}N uniformly labeled and ^{14}N selectively and inversely labeled IL1βs. We expressed them in *E. coli* BL21(*DE3*) grown in M9 minimal media that contained $^{15}NH_4Cl$ as the sole nitrogen source with or without supplementation of specific ^{14}N-labeled amino acids at concentrations of 100 mg/L each. The ^{14}N amino acids were added to the cultures approximately 30–60 min prior to IPTG-induced protein expression. Because *E. coli* BL21(*DE3*) was a competent host strain, no special auxotrophic host strains were needed. The simplified fermentation protocol is summarized in Table 1.

We carefully selected combinations of amino acids to be inversely labeled. In detail, these amino acids are the final products of isolated metabolic pathways or are synthesized in metabolically related pathways. The expression of ^{15}N uniformly labeled IL1βs was possible in *E. coli* BL21(*DE3*), which produces large amounts of recombinant proteins in the minimal media [19]. We expressed the labeled IL1β in BL21(*DE3*) cells from plasmid DNA under the control of the p*tac* promoter. Next, expression for other ^{14}N inversely labeled proteins was examined. We chose three metabolically isolated amino acids (Arg, Lys, and Ala) and four combinations of amino acids (Phe/Tyr, Ile/Val/Leu, His/Trp, and Gly/Ser/Cys) to label IL1β. Note that the biosynthetic pathways for His and Trp are metabolically distinct from each other, and hence a protein need

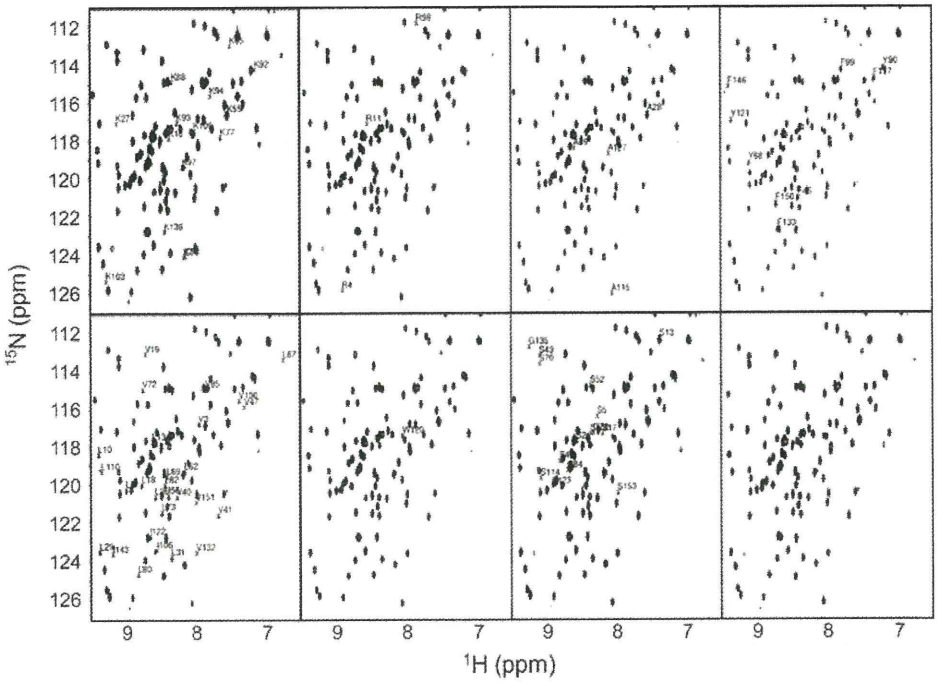

Fig. 1 Expanded regions of the ^{14}N selectively and inversely labeled IL1βs' ^{1}H–^{15}N HSQC spectra. The inversely labeled proteins, named according to the ^{14}N-labeled residues that they contain, are **a** Lys^{-}, **b** Arg^{-}, **c** Ala^{-}, **d** (Phe/Tyr)$^{-}$, **e** (Ile/Val/Leu)$^{-}$, **f** (His/Trp)$^{-}$, and **g** (Gly/Ser/Cys)$^{-}$. The control spectrum, which is that of fully labeled IL1β, is shown in **h**. The full range expansion spectrum of (**h**) with assignments is shown in Supplementary Fig. 1. The eliminated *cross peaks* are marked by *crosses* and residue names

Table 2 Cross peak identification in the ^{1}H–^{15}N HSQC of IL1β inversely labeled with the selected ^{14}N amino acid(s)

Amino acid(s)	Number of expected cross peaks	Number of eliminated cross peaks	Number of overlapping cross peaks[a]
K	15	14	1
R	3	3	0
A	4	3	1
I/V/L	5/11/15	31	0
H/W	1/1	2	0
G/S/C	8/14/2	23	1
F/Y	9/4	9[b]	0

[a] The positions of these peak(s) overlap with other peaks. It was not possible to identify them as eliminated peaks by manual inspection

[b] Four of these peaks were barely visible but could be identified as phenylalanine or tyrosine cross peaks

not be labeled with both amino acids at once. Nevertheless, in this study, we used these two amino acids simultaneously because His and Trp are usually present in low concentrations in the protein. We could identify each residue type by its absent or weakened cross peak in the reference ^{1}H–^{15}N HSQC spectrum (Fig. 1). The targeted cross peaks' intensities decreased by greater than 95% for all spectra, while the method did not affect the other cross peak intensities (Table 2). The incorporation ratio of ^{14}N inversely labeled amino acids by each method was quantified (Supplementary Table 1). Thus, our method can provide a large chunk of the missing information, specifically, the residue-type information associated with amide cross peaks, even if many ^{13}Cβ chemical shifts in the HNCACB and CBCA (CO)NH spectra are absent.

We also employed our method for de novo assignment of several protein NMR spectra, for example, those of the PX domain from human p47^{phox} NADPH oxidase (PX-p47^{phox}, residues 1–128, 130 total residues) [20], the DNA binding domain of the *Drosophila melanogaster gcm* transcriptional factor (*gcm*-DBD, 158 residues) [21, 22], the TIR domain from MyD88 (residues 148–296, 149 residues) [23, 24], and the second LOV domain (LOV2) from *Arabidopsis thaliana* phototropin 2 (LOV2^{phot2}, residues 363–500, 140 total residues; manuscript in preparation). In all cases, 3D-NMR data sets were analyzed using only spectra recorded with fully ^{13}C/^{15}N-labeled protein samples prepared at concentrations of 0.3–0.6 mM—perdeuterated samples were not needed. For PX-p47^{phox}, all observable amide signals were assigned when the data sets for PX-p47^{phox} that had been selectively and inversely labeled with arginines and lysines (designated Arg^{-} and Lys^{-}, respectively) were used to eliminate ambiguities. For *gcm*-DBD, more than 95% of the backbone signals were unambiguously assigned by incorporating the information contained in the ^{1}H–^{15}N HSQC spectra of Arg^{-} and Lys^{-} samples during an iterative computer assignment followed by manual examination. The assignment results were further confirmed using the data obtained from the ^{1}H–^{15}N HSQC spectra of (Phe/Tyr)$^{-}$ and (His/Trp)$^{-}$ samples. Assignment of the gcm-DBD resonances proved to be our most difficult test case because of poor signal-to-noise

ratios found for many of the 3D NMR spectra. Only 60% of the intra-residue CO, 30% of the inter-residue Cβ, and 10% of the intra-residue Cβ cross peaks were observed in the HN (CA)CO, CBCA(CO)NH, and HNCACB spectra, respectively (Fig. 2). For LOV2^{phot2}, more than 95% of its backbone signals were unambiguously assigned when the data sets obtained from the ^{1}H–^{15}N HSQC spectra of (Ile/Val/Leu)$^{-}$ and (Phe/Tyr)$^{-}$ were included in the assignment procedure. Note that these examples included not only a T7 promoter-driven pET-based expression system, which was originally described by Shortle [13], but also a *tac*-promoter-based pGEX GST-fusion protein expression system. We further confirmed the expression of the inversely labeled sample with a pMal MBP-fusion system as well. In essence, there is no difference between these expression systems.

One possible problem associated with this combinatorial labeling method concerns potential ambiguity of assignment(s) when an eliminated cross peak overlaps another cross peak of any residue type. Usually, it will be easy to classify resonance(s) according to residue type without ambiguity when the ^{1}H–^{15}N HSQC spectrum of the inversely labeled sample and the reference spectrum are compared. If there are not many unassigned backbone resonances to begin with, a visual examination of the two spectra should suffice. However, it often becomes difficult to begin to assign backbone resonances because many of the necessary inter- and intra-residue cross peaks, especially those involving Cβs and COs, are missing within a data set. These NMR cross peaks are needed to make connectivities between two successive residues; however, the cross peaks between backbone amides and the intra-residual Cβs and COs tend to vanish. This is because the two triple-resonance experiments HNCACB and HN(CA)CO are remarkably insensitive compared to CBCA(CO)NH and HNCO, respectively, in cases of non-ideal samples. When the aforementioned difficulty occurs, it may be resolved if an automated assignment program (e.g., AUTOASSIGN [25], MAPPER [26], MARS [27], KUJIRA [28], or Olivia [29]) is used in conjunction with the additional data acquired from inverse-labeling experiments. For the experimental work reported herein, we used a program developed in-house that relies on a simple exhaustive search algorithm for residue assignment. The data set inputted into these assignment programs usually consists of peak lists obtained from six 3D NMR experiments: HNCA, HNCO, HN(CO)CA, HN(CA)CO, CBCA(CO)NH, and HNCACB. The residue-type information obtained from ^{1}H–^{15}N HSQC spectra of inversely labeled proteins would be added to these input data. Using our simplified method, selected combinations of amide signal elimination data were added in a stepwise manner according to the completion of the assignment. During the assignment procedure, the cross peak data generated from a 3D ^{15}N-edited NOESY spectrum should not be used as additional input because this data can be used later to

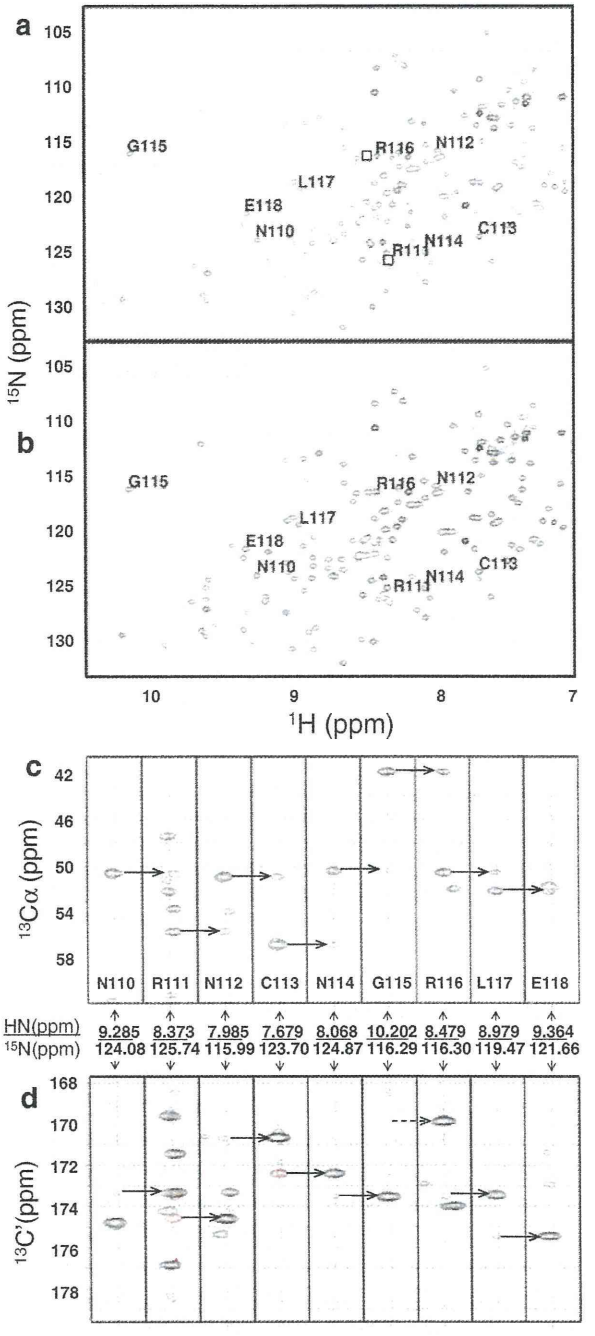

Fig. 2 An example of the backbone assignment of GCM-DBD using inversely labeled protein. Expanded regions of the of the ^{14}Arg-labeled and fully labeled references and GCM-DBD's ^{1}H–^{15}N HSQC spectra are shown in **a** and **b**, respectively. The eliminated signals for R111 and R116 are indicated as *open squares*. Selected *strip plots* of **c** HNCA and **d** HNCO (*black*)/HNCACO (*red*) spectra overlays corresponding to the residues N110 to E118 are shown. Sequential connectivities between strips are indicated by *arrows*

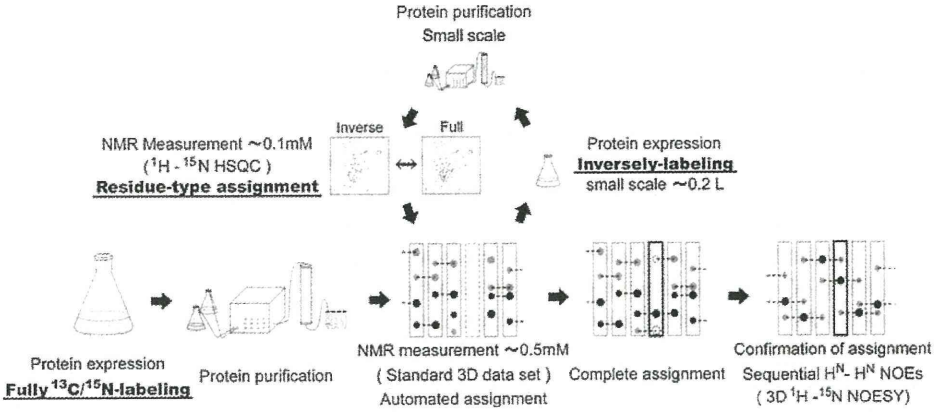

Fig. 3 Scheme for the iterative procedure for assignment of backbone resonances using inversely labeled protein

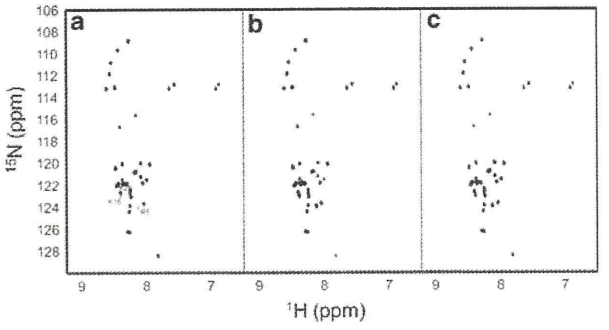

Fig. 4 Expanded regions of the ^{14}N selectively and inversely labeled Aβ(1–40)s' ^{1}H–^{15}N SOFAST-HMQC spectra. The inversely labeled proteins, named according to the ^{14}N-labeled residues that they contain, are **a** (Arg/Lys)$^{-}$ and **b** (Ile/Val/Leu)$^{-}$. The control spectrum, which is that of fully labeled Aβ(1–40), is shown in **c**. The inversely labeled *cross peaks* are marked by crosses and residue names

eliminate any assignment ambiguities. Many (e.g., 10–30) sequential assignment possibilities may be generated by a semi-automatic program, but often the correct assignment can be confirmed using the ^{15}N-edited NOESY connectivity data. Finally, backbone resonance assignments can be further confirmed by manually tracing the connectivity of the sequential NOEs. Interpretation of the spectra of non-ideal protein samples should benefit from the use of the strategy outlined above because NOESY experiments usually have greater signal-to-noise ratios compared with HNCACB and HN(CA)CO experiments. Our NMR assignment procedure using inversely labeled protein is summarized in Fig. 3.

Another limitation of this method is the harvesting time of the bacteria after the addition of the ^{14}N amino acids. For example, some aggregation-prone proteins are available in their soluble form only when the proteins are expressed at a low temperature. For assessing this problem, we prepared (Arg/Lys)$^{-}$ and (Ile/Leu/Val)$^{-}$ Aβ(1–40) peptide (40 residues, Fig. 4) as a good example of an aggregation-prone protein sample. The sample was prepared by expressing a maltose binding protein fusion form of Aβ(1–40) at 20°C for 18 h using a pMAL-c2-based plasmid, followed by tag removal by TEV protease and purification by HPLC. While one lysine and two arginine peaks were eliminated in the spectrum of (Arg/Lys)$^{-}$, no inverse-labeling effect was observed in the spectrum of (Ile/Leu/Val)$^{-}$ because of metabolic scrambling. Therefore, long culture times such as 18 h after IPTG induction should be avoided.

The merit of our simplified method compared with existing labeling methods (e.g., residue-specific ^{15}N-labeling methods [30, 31] and a side chain/residue-specific ^{12}C-inverse-labeling method, where other carbons are ^{13}C-labeled [32]) is in the preparation of proteins with specific, metabolically related amino acids that are ^{14}N inversely labeled in combination. Although our method requires two to seven additional protein samples with different ^{14}N inversely labeled amino acids, the required concentration of each sample is less than 0.1 mM, which suffices for a single experiment of ^{1}H–^{15}N HSQC. Such samples are routinely prepared economically with a small scale fermentation (i.e., 0.2 L). In addition, inverse-labeling is routinely achieved by expressing the target protein in *E. coli* strain BL21(*DE3*) with the normal glucose-based M9 minimal medium—auxotrophic host strains, cell-free protein expression systems, or specialized glycerol-based MOPS medium are not required. This implies that one can skip the step for optimizing the conditions of individual experiments to prepare the additional inversely labeled samples. Thus, we found our method experimentally advantageous because it can be used additively after the standard NMR assignment experiments using the fully labeled sample, even when the quality of some of the 3D NMR spectra were found to be insufficient to complete the full assignment. Finally, this approach is not necessarily limited by sample viscosity (suitable for highly viscous samples) or protein locality (compatible with an in-cell NMR approach with *E. coli*) [33].

Conclusion

Our simplified combinatorial ^{14}N selective- and inverse-labeling method has enormous potential for NMR-based structural biology studies of proteins, because the strategy should be widely applicable for proteins with non-ideal properties at any stage of the research project. The information from the spectra of ^{14}N selectively and inversely labeled proteins could be applied to compensate for an otherwise incomplete 3D NMR data set in a stepwise manner. The method reduces the ambiguity often present during the initial sequential assignment trials and can confirm the tentative assignments derived from 3D-NMR spectra of marginal quality because the specific residue type of a cross peak can be readily identified as a result of the elimination of its cross peak in the ^{1}H–^{15}N HSQC spectrum.

Acknowledgments We thank the people who helped to confirm the versatility of the method by applying it to each of the proteins: Dr. M. Shimizu (*gcm*-DBD), Mr. M. Itoh (LOV2^{phot}), Dr. H. Tochio and Dr. H. Ohnishi (TIR-MyD88). We also thank Dr. D. Hamada for providing an Aβ expression system.

References

1. Schanda P, Kupce E, Brutscher B (2005) SOFAST-HMQC experiments for recording two-dimensional heteronuclear correlation spectra of proteins within a few seconds. J Biomol NMR 33:199–211
2. Pervushin K, Riek R, Wider G, Wuthrich K (1997) Attenuated T2 relaxation by mutual cancellation of dipole–dipole coupling and chemical shift anisotropy indicates an avenue to NMR structures of very large biological macromolecules in solution. Proc Natl Acad Sci USA 94:12366–12371
3. Cavanagh J, Fairbrother WJ, Palmer AG III, Skelton NJ (2007) Protein NMR spectroscopy: principles and practice, 2nd edn. Academic Press, San Diego, CA
4. Ikura M, Kay LE, Bax A (1990) A novel approach for sequential assignment of ^{1}H, ^{13}C and ^{15}N spectra of proteins: heteronuclear triple-resonance three-dimensional NMR spectroscopy. Appl Calmodulin Biochem 29:4659–4667
5. Yamazaki T, Lee W, Arrowsmith CH, Muhandiram DR, Kay LE (1994) A Suite of Triple Resonance NMR Experiments for the Backbone Assignment of ^{15}N, ^{13}C, ^{2}H Labeled Proteins with High Sensitivity. J Am Chem Soc 116:11655–11666
6. Schanda P, Van Melckebeke H, Brutscher B (2006) Speeding up three-dimensional protein NMR experiments to a few minutes. J Am Chem Soc 128:9042–9043
7. Barna JCJ, Laue ED, Mayger MR, Skilling J, Worrall SJP (1987) Exponential sampling an alternative method for sampling in two-dimensional NMR experiments. J Magn Reson 73:69–77
8. Schmieder P, Stern AS, Wagner G, Hoch JC (1994) Improved resolution in triple-resonance spectra by nonlinear sampling in the constant-time domain. J Biomol NMR 4:483–490
9. Rovnyak D, Frueh DP, Sastry M, Sun ZY, Stern AS, Hoch JC, Wagner G (2004) Accelerated acquisition of high resolution triple-resonance spectra using non-uniform sampling and maximum entropy reconstruction. J Magn Reson 170:15–21
10. Kainosho M, Tsuji T (1982) Assignment of the three methionyl carbonyl carbon resonances in Streptomyces subtilisin inhibitor by a carbon-13 and nitrogen-15 double-labeling technique. A new strategy for structural studies of proteins in solution. Biochemistry 21:6273–6279
11. Lee KM, Androphy EJ, Baleja JD (1995) A novel method for selective isotope labeling of bacterially expressed proteins. J Biomol NMR 5:93–96
12. Morita EH, Shimizu M, Ogasawara T, Endo Y, Tanaka R, Kohno T (2004) A novel way of amino acid-specific assignment in 1H–15N HSQC spectra with a wheat germ cell-free protein synthesis system. J Biomol NMR 30:37–45
13. Shortle D (1994) Assignment of amino acid type in 1H–15N correlation spectra by labeling with 14N-amino acids. J Magn Reson B 105:88–90
14. Neidhardt FC, Bloch PL, Smith DF (1974) Culture medium for enterobacteria. J Bacteriol 119:736–747
15. Kikumoto Y, Hong YM, Nishida T, Nakai S, Masui Y, Hirai Y (1987) Purification and characterization of recombinant human interleukin-1 beta produced in *Escherichia coli*. Biochem Biophys Res Commun 147:315–321
16. Delaglio F, Grzesiek S, Vuister GW, Zhu G, Pfeifer J, Bax A (1995) NMRPipe:a multidimensional spectral processing system based on UNIX pipes. J Biomol NMR 6:277–293
17. Tate S, Kikumoto Y, Ichikawa S, Kaneko M, Masui Y, Kamogashira T, Ouchi M, Takahashi S, Inagaki F (1992) Stable isotope aided nuclear magnetic resonance study to investigate the receptor-binding site of human interleukin 1 beta. Biochemistry 31:2435–2442
18. Hou L, Zagorski MG (2006) NMR reveals anomalous copper (II) binding to the amyloid Abeta peptide of Alzheimer's disease. J Am Chem Soc 128:9260–9261
19. Studier FW, Moffatt BA (1986) Use of bacteriophage T7 RNA polymerase to direct selective high-level expression of cloned genes. J Mol Biol 189:113–130
20. Hiroaki H, Ago T, Ito T, Sumimoto H, Kohda D (2001) Solution structure of the PX domain a target of the SH3 domain. Nat Struct Biol 8:526–530
21. Shimizu M, Hiroaki H, Kohda D, Morita EH, Hotta S, Morikawa K (2003) ^{1}H, ^{13}C and ^{15}N backbone resonance assignments of the N-terminal domain of *Drosophila* GCM protein. J Biomol NMR 26:277–278
22. Shimizu M, Hiroaki H, Kohda D, Hosoya T, Akiyama-Oda Y, Hotta Y, Morita EH, Morikawa K (2003) NMR and ICP spectroscopic analysis of the DNA-binding domain of the Drosophila GCM protein reveals a novel Zn2+ -binding motif. Protein Eng 16:247–254
23. Ohnishi H, Tochio H, Kato Z, Kimura T, Hiroaki H, Kondo N, Shirakawa M (2010) ^{1}H, ^{13}C and ^{15}N resonance assignment of the TIR domain of human MyD88. Biomol NMR Assign 4:123–125
24. Ohnishi H, Tochio H, Kato Z, Orii KE, Li A, Kimura T, Hiroaki H, Kondo N, Shirakawa M (2009) Structural basis for the multiple interactions of the MyD88 TIR domain in TLR4 signaling. Proc Natl Acad Sci USA 106:10260–10265
25. Zimmerman DE, Kulikowski CA, Huang Y, Feng W, Tashiro M, Shimotakahara S, Chien C, Powers R, Montelione GT (1997) Automated analysis of protein NMR assignments using methods from artificial intelligence. J Mol Biol 269:592–610
26. Guntert P, Salzmann M, Braun D, Wuthrich K (2000) Sequence-specific NMR assignment of proteins by global fragment mapping with the program MAPPER. J Biomol NMR 18:129–137
27. Jung YS, Zweckstetter M (2004) Mars–robust automatic backbone assignment of proteins. J Biomol NMR 30:11–23

28. Kobayashi N, Iwahara J, Koshiba S, Tomizawa T, Tochio N, Guntert P, Kigawa T, Yokoyama S (2007) KUJIRA, a package of integrated modules for systematic and interactive analysis of NMR data directed to high-throughput NMR structure studies. J Biomol NMR 39:31–52
29. Yokochi M, Sekiguchi S, Inagaki F (2010) Olivia. Hokkaido University, Sapporo
30. Ou HD, Lai HC, Serber Z, Dotsch V (2001) Efficient identification of amino acid types for fast protein backbone assignments. J Biomol NMR 21:269–273
31. Strauss A, Bitsch F, Cutting B, Fendrich G, Graff P, Liebetanz J, Zurini M, Jahnke W (2003) Amino-acid-type selective isotope labeling of proteins expressed in Baculovirus-infected insect cells useful for NMR studies. J Biomol NMR 26:367–372
32. Vuister GW, Kim SJ, Wu C, Bax A (1994) 2D and 3D NMR study of phenylalanine residues in proteins by reverse isotopic labeling. J Am Chem Soc 116:9206–9210
33. Serber Z, Dotsch V (2001) In-cell NMR spectroscopy. Biochemistry 40:14317–14323

Published online 8 November 2011 *Nucleic Acids Research, 2012, Vol. 40, Database issue* ***D507–D511***
doi:10.1093/nar/gkr884

IDEAL: Intrinsically Disordered proteins with Extensive Annotations and Literature

Satoshi Fukuchi[1,*], Shigetaka Sakamoto[2], Yukiko Nobe[3], Seiko D. Murakami[3], Takayuki Amemiya[3], Kazuo Hosoda[1], Ryotaro Koike[3], Hidekazu Hiroaki[4] and Motonori Ota[3,*]

[1]Faculty of Engineering, Maebashi Institute of Technology, Maebashi, Gunma 371-0816, [2]HOLONICS Corporation, Numazu, Shizuoka 411-0803, [3]Graduate School of Information Sciences and [4]Graduate School of Science, Nagoya University, Nagoya 464-8601, Japan

Received August 11, 2011; Revised September 27, 2011; Accepted September 30, 2011

ABSTRACT

IDEAL, Intrinsically Disordered proteins with Extensive Annotations and Literature (http://www.ideal.force.cs.is.nagoya-u.ac.jp/IDEAL/), is a collection of knowledge on experimentally verified intrinsically disordered proteins. IDEAL contains manual annotations by curators on intrinsically disordered regions, interaction regions to other molecules, post-translational modification sites, references and structural domain assignments. In particular, IDEAL explicitly describes protean segments that can be transformed from a disordered state to an ordered state. Since in most cases they can act as molecular recognition elements upon binding of partner proteins, IDEAL provides a data resource for functional regions of intrinsically disordered proteins. The information in IDEAL is provided on a user-friendly graphical view and in a computer-friendly XML format.

INTRODUCTION

The discovery of intrinsically disordered proteins (IDPs) has brought about a paradigm change in structural biology (1,2). Although proteins were believed to adopt unique 3D structures to function, IDPs do not, by themselves, assume any stable 3D structure under physiological conditions, and yet they participate in crucial biological processes such as signal transduction and transcription control (3–5). Some proteins contain long intrinsically disordered regions (IDRs) while others are fully disordered. In contrast to the long studied 3D structures of proteins, investigations on IDPs started only about 10 years ago and, as yet, knowledge of IDPs is not well collected and integrated. Although the first database of IDPs, Disprot (6), has more than 600 well-annotated entries, this number is much smaller than the over 70 thousands entries in the Protein Data Bank (PDB) (7). Considering that the protein 3D structural databases such as PDB, SCOP (Structural Classification of Proteins) (8) and CATH (9), have played important roles in deepening our understanding of the nature of protein structures and functions, the development of IDP databases are essential to the progress of IDP research.

We have developed a database, IDEAL (IDPs with Extensive Annotations and Literature) in which experimentally verified IDRs are collected. In the database construction process, we paid special attention to the functional regions in IDRs, for example, regions that interact with other molecules and post-translational modification sites. In particular, we have extensively curated IDRs that adopt unique 3D structures when they bind to other molecules by the 'coupled folding and binding' process (10–16). We have called these IDRs the protean segments (ProS). The information in IDEAL is provided on a user-friendly web-interface and in computer-friendly XML files.

CONTENTS OF IDEAL

Summary of the annotation process

We used the UniProt amino acid sequence (17) as the reference, and marked structural and functional features along the sequences. A unique serial identifier, IID (IDEAL Identification), was assigned to each protein in IDEAL, starting with IID0001 for human proteins, IID5001 for other eukaryotic proteins and IID9001 for all other proteins including virus proteins. Ordered and disordered regions were annotated as follows: First, ordered regions were obtained from the structural regions atomically detailed in the PDB. Then, disordered

*To whom correspondence should be addressed. Tel/Fax: +81-27-265-7376; Email: sfukuchi@maebashi-it.ac.jp
Correspondence may also be addressed to Motonori Ota. Tel/Fax: +81-52-789-4782; Email: mota@is.nagoya-u.ac.jp

regions were located by careful assessment of PDB coordinates and by reading the literature. After identifying the ordered and disordered regions, the ProSs were manually determined. Finally, miscellaneous information, such as binding sites and post-translational modifications, was derived mainly from UniProt annotations, and structural domains were assigned by homology searches.

Proteins stored in IDEAL

As a starting point for the annotation, we chose UniProt human nuclear proteins with PDB structures (712 proteins), because eukaryotic nuclear proteins are known to contain long IDRs (18,19). We have annotated more than 120 human nuclear proteins. Out of them, the overlap with DisProt is only one-third at most, indicating IDEAL and DisProt complement each other. Most of the PDB structures for these proteins are hetero-oligomers in which the protein was associated with its binding partners. Annotations for these partner proteins are also in IDEAL, regardless of the source organism or the presence or absence of IDRs.

Ordered/disordered regions

The most important part of the IDEAL annotation is to identify the ordered and disordered amino acid segments. Ordered regions can be assigned by referring to the PDB. It is not straightforward to identify the disordered regions. In IDEAL, disordered regions are judged using several criteria; (i) missing residues in the X-ray structures, (ii) regions that interfere with protein crystallization in X-ray experiments, (iii) regions that fluctuate greatly in ensemble number of NMR model structures and (iv) regions that have been shown to be flexible in experiments using NMR, CD and other methods, and that have no corresponding structures in PDB. Of the four categories, (i) can be automatically obtained from PDB. Regions that were identified using the other three categories could only be judged manually. Although fluctuating regions in category (iii) could be found automatically by comparing the PDB coordinates of a group of models, the regions were only accepted as IDRs after curators confirmed the fluctuations by examining the corresponding literature. Category (iv) requires the most laborious procedure to be obtained, but provides variable information. Curators conduct manual literature searches to obtain such information as much as possible.

Protean segment

One of the reasons why IDPs have drawn so much attention is the discovery of the phenomenon known as coupled folding and binding in which a short flexible segment binds to its binding partner by forming a specific structure which acts as the molecular recognition element (10–16). In IDEAL, we explicitly annotated this short flexible region as ProS when both unstructured and structured information is available for the region. We defined two categories for ProS, verified ProS and possible ProS. A verified ProS is a sequence for which there is evidence of both a disordered isolated state and an ordered binding state. A possible ProS is a sequence for which there is only evidence of an ordered binding state, but circumstantial evidence suggests that the sequence is disordered in the isolated state. A possible ProS is, for example, a sequence from a protein whose homolog contains a verified ProS in the corresponding position. Another example would be the one in which the binding partner of a possible ProS binds a verified ProS using the same interface.

Sequences involved in coupled folding and binding have been addressed in several ways, for example, molecular recognition features (MoRFs) (20) and eukaryotic linear motifs (ELMs) (21) have been studied. Although ProS, MoRF and ELM are similar concepts, MoRF has a length limitation of 70 residues and an ELM should have a motif that can be described in a regular expression. On the other hand, the definition of ProS depends only on evidence of a disorder-order transition. Although most ProSs bind to a partner protein, by its definition, ProS can include IDRs whose structures are induced upon binding to small ligands. ProSs do not necessarily assume secondary structures in the binding state, and long IDRs or IDRs without a motif can also be ProSs. Some relatively long IDRs, such as p27Kip1 (PDB:1jsu) and Tcf3 (PDB:1g3j), can transform into ordered states (22). ProSs can also cover these IDRs.

MISCELLANEOUS INFORMATION

We integrate the miscellaneous information from UniProt, namely, regions interacting with other molecules, motifs and post-translational modifications. During the annotation process, the curators find interaction sites, sequence motifs or other information that has not been described in UniProt, the new information is included in IDEAL. IDEAL also provides SCOP (version 1.75) and Pfam (23) (version 24.0) domain assignments using reverse PSI-Blast (24) and HMMer (25). Note that ordered regions assigned in the order/disorder annotation process are experimentally verified ordered regions, while the structural domain assignments were done using homology searches.

USING IDEAL

Browse and search entries

'The list' on the top page of IDEAL provides an easy way to access any of the entries in IDEAL. The list enumerates all entries in IDEAL, where IID, protein name, organism, total sequence length and the presence/absence of ProS are tabulated. IDEAL also provides a search tool, which always appears in the blue bar at the top of each page ([1] in Figure 1). Users can choose from 'Full text', 'UniProt accession', 'Protein Name' and 'PDB id' categories, and enter some words or an ID into the input field. The BLAST search is available through the 'BLAST search' link button, and the user can input an amino acid sequence to find homologs in the IDEAL entries.

Representation of each entry

IDEAL provides a user-friendly web interface for each entry. An example, a page for catenin β-1, is shown

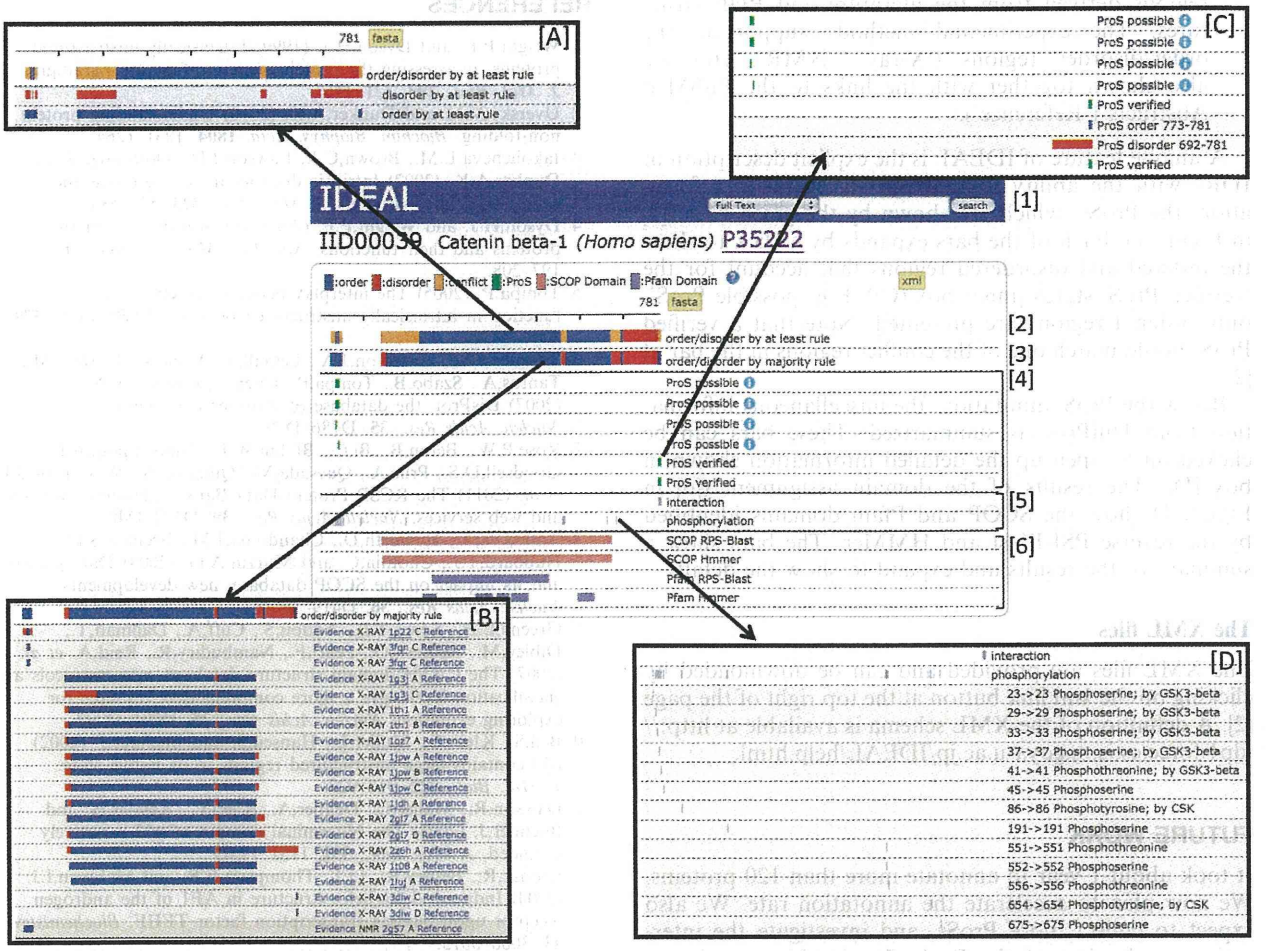

Figure 1. IDEAL annotation for catenin β-1. The identifier, IID, protein name, source organism and the link to UniProt are shown below the blue bar which contains the search tool ([1]). Bars [2] and [3] show a summary of the annotated regions using the 'at least rule' and 'majority rule' criteria. Ordered, disordered and conflict regions are colored in blue, red and orange, respectively. The inner boxes (**A**) and (**B**) show a detailed breakdown of the regions in [2] and [3], respectively. These diagrams appear by clicking on the bars. The inner box (**B**) provides the link to the experimental evidence and the technique supporting the order/disorder regions. ProSs are represented in section [4] as the green bars, which can be expanded by clicking to reveal the inner box (**C**). Interaction sites and post-translational modification sites are shown in section [5]. These bars expand to show the detailed information in the inner box (**D**). The bars in section [6] show the domain assignments.

in Figure 1. The annotated regions are presented in a bar diagram to help make annotations intuitively understandable. Two color bars at the top ([2] and [3] in Figure 1) summarize ordered/disordered information in the distinctive ways shown. A protein may have multiple PDB entries and other information without accompanying PDB entries from different experimental techniques such as CD, H/D exchange, etc. Because IDEAL contains all PDB entries together with the other structural information associated with a query protein, all the associated information are not necessarily consistent due to different experimental conditions and other reasons. To summarize these diverse situations, IDEAL uses two representations:

(1) The bar [2] in Figure 1 shows the summary of ordered/disordered regions by the 'at least rule'. Here an ordered (blue bar) or disordered (red bar) site is shown if the site has at least one ordered or one disordered annotation. When a single site has both an ordered and a disordered annotation, the site is in 'conflict' (orange bar). The inner box [A] in Figure 1, opened by clicking the bar, shows the detailed breakdown of the annotations. The first and the second bars correspond to the at least ordered regions, and the at least disordered regions, respectively. All of the data sources supporting each of order/disorder regions can be presented by cricking the 'majority rule' bar explained below.

(2) The bar at [3] shows the summary of ordered/disordered regions by the 'majority rule', in which majority decision is employed to show the annotation. The inner-box (B), opened by clicking the bar, shows all the evidence of annotations used to in the majority vote. They include ordered and disordered

regions derived from the literature and PDB structures. The experimental methods supporting the order/disorder regions ('X-ray', 'NMR', etc) are also shown together with the links to the PubMed Abstracts ('Reference').

A unique feature of IDEAL is the explicit description of IDRs with the ability to undergo structural transformation, the ProSs, which are shown by the green bars ([4] in Figure 1). Each of the bars expands by a click to show the ordered and disordered regions that account for the 'verified ProS' status [inner box (C)]. For 'possible ProS', only ordered regions are presented. Note that a verified ProS should match one of the conflict regions in the bar at [2].

Below the ProS annotation, the miscellaneous information from UniProt, is summarized. These bars can be clicked on to open up the detailed information shown in box [D]. The results of the domain assignment ([6] in Figure 1) show the SCOP and Pfam domains identified by the reverse PSI-Blast and HMMer. The bars show a summary of the results and expand to show the details.

The XML files

The XML files are provided and can be downloaded by clicking on the xml link button at the top right of the page [2]. A definition of the XML schema is available at http://idp1.force.cs.is.nagoya-u.ac.jp/IDEAL/help.html.

FUTURE WORK

It took about 1 year to annotate more than 120 proteins. We now plan to accelerate the annotation rate. We also expect to collect more ProSs, and investigate the interaction mechanism of the ProS. To do this, we aim to develop an interface showing the binding partner proteins associated with ProSs and to illustrate their interaction networks. As in any databases, updating the contents is a key issue. We will address this by developing an update system to keep information in IDEAL as current as possible.

ACKNOWLEDGEMENTS

The authors thank Masahito Umezaki and Tomoko Sato for their contributions at the beginning of the project. The authors also thank Keiichi Homma for his valuable suggestions.

FUNDING

Grant-in-Aid for Scientific Research on Innovative Areas, 'Target recognition and expression mechanism of intrinsically disordered proteins' from the Ministry of Education, Culture, Sports, Science, and Technology (MEXT) of Japan. Funding for open access charge: MEXT.

Conflict of interest statement. None declared.

REFERENCES

1. Wright,P.E. and Dyson,H.J. (1999) Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm. *J. Mol. Biol.*, **293**, 321–331.
2. Uversky,V.N. and Dunker,A.K. (2010) Understanding protein non-folding. *Biochim. Biophys. Acta*, **1804**, 1231–1264.
3. Iakoucheva,L.M., Brown,C.J., Lawson,J.D., Obradovic,Z. and Dunker,A.K. (2002) Intrinsic disorder in cell-signaling and cancer-associated proteins. *J. Mol. Biol.*, **323**, 573–584.
4. Dyson,H.J. and Wright,P.E. (2005) Intrinsically unstructured proteins and their functions. *Nat. Rev. Mol. Cell. Biol.*, **6**, 197–208.
5. Tompa,P. (2005) The interplay between structure and function in intrinsically unstructured proteins. *FEBS Lett.*, **579**, 3346–3354.
6. Sickmeier,M., Hamilton,J.A., LeGall,T., Vacic,V., Cortese,M.S., Tantos,A., Szabo,B., Tompa,P., Chen,J., Uversky,V.N. *et al.* (2007) DisProt: the database of disordered proteins. *Nucleic Acids Res.*, **35**, D786–D793.
7. Rose,P.W., Beran,B., Bi,C., Bluhm,W.F., Dimitropoulos,D., Goodsell,D.S., Prlic,A., Quesada,M., Quinn,G.B., Westbrook,J.D. *et al.* (2011) The RCSB Protein Data Bank: redesigned web site and web services. *Nucleic Acids Res.*, **39**, D392–D401.
8. Andreeva,A., Howorth,D., Chandonia,J.M., Brenner,S.E., Hubbard,T.J., Chothia,C. and Murzin,A.G. (2008) Data growth and its impact on the SCOP database: new developments. *Nucleic Acids Res.*, **36**, D419–D425.
9. Greene,L.H., Lewis,T.E., Addou,S., Cuff,A., Dallman,T., Dibley,M., Redfern,O., Pearl,F., Nambudiry,R., Reid,A. *et al.* (2007) The CATH domain structure database: new protocols and classification levels give a more comprehensive resource for exploring evolution. *Nucleic Acids Res.*, **35**, D291–D297.
10. Bell,S., Klein,C., Muller,L., Hansen,S. and Buchner,J. (2002) p53 contains large unstructured regions in its native state. *J. Mol. Biol.*, **322**, 917–927.
11. Dawson,R., Muller,L., Dehner,A., Klein,C., Kessler,H. and Buchner,J. (2003) The N-terminal domain of p53 is natively unfolded. *J. Mol. Biol.*, **332**, 1131–1141.
12. Kumar,R., Betney,R., Li,J., Thompson,E.B. and McEwan,I.J. (2004) Induced alpha-helix structure in AF1 of the androgen receptor upon binding transcription factor TFIIF. *Biochemistry*, **43**, 3008–3013.
13. Lee,H., Mok,K.H., Muhandiram,R., Park,K.H., Suk,J.E., Kim,D.H., Chang,J., Sung,Y.C., Choi,K.Y. and Han,K.H. (2000) Local structural elements in the mostly unstructured transcriptional activation domain of human p53. *J. Biol. Chem.*, **275**, 29426–29432.
14. Nagadoi,A., Nakazawa,K., Uda,H., Okuno,K., Maekawa,T., Ishii,S. and Nishimura,Y. (1999) Solution structure of the transactivation domain of ATF-2 comprising a zinc finger-like subdomain and a flexible subdomain. *J. Mol. Biol.*, **287**, 593–607.
15. Receveur-Brechot,V., Bourhis,J.M., Uversky,V.N., Canard,B. and Longhi,S. (2006) Assessing protein disorder and induced folding. *Proteins*, **62**, 24–45.
16. Rustandi,R.R., Baldisseri,D.M. and Weber,D.J. (2000) Structure of the negative regulatory domain of p53 bound to S100B(betabeta). *Nat. Struct. Biol.*, **7**, 570–574.
17. UniProt_Consortium. (2010) The Universal Protein Resource (UniProt) in 2010. *Nucleic Acids Res.*, **38**, D142–D148.
18. Fukuchi,S., Hosoda,K., Homma,K., Gojobori,T. and Nishikawa,K. (2011) Binary classification of protein molecules into intrinsically disordered and ordered segments. *BMC Struct. Biol.*, **11**, 29.
19. Ward,J.J., Sodhi,J.S., McGuffin,L.J., Buxton,B.F. and Jones,D.T. (2004) Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. *J. Mol. Biol.*, **337**, 635–645.
20. Mohan,A., Oldfield,C.J., Radivojac,P., Vacic,V., Cortese,M.S., Dunker,A.K. and Uversky,V.N. (2006) Analysis of molecular recognition features (MoRFs). *J. Mol. Biol.*, **362**, 1043–1059.
21. Gould,C.M., Diella,F., Via,A., Puntervoll,P., Gemund,C., Chabanis-Davidson,S., Michael,S., Sayadi,A., Bryne,J.C., Chica,C.

et al. (2010) ELM: the status of the 2010 eukaryotic linear motif resource. *Nucleic Acids Res.*, **38**, D167–D180.
22. Tompa,P., Fuxreiter,M., Oldfield,C.J., Simon,I., Dunker,A.K. and Uversky,V.N. (2009) Close encounters of the third kind: disordered domains and the interactions of proteins. *Bioessays*, **31**, 328–335.
23. Finn,R.D., Mistry,J., Tate,J., Coggill,P., Heger,A., Pollington,J.E., Gavin,O.L., Gunasekaran,P., Ceric,G., Forslund,K. *et al.* (2010) The Pfam protein families database. *Nucleic Acids Res.*, **38**, D211–D222.
24. Marchler-Bauer,A., Lu,S., Anderson,J.B., Chitsaz,F., Derbyshire,M.K., DeWeese-Scott,C., Fong,J.H., Geer,L.Y., Geer,R.C., Gonzales,N.R. *et al.* (2011) CDD: a Conserved Domain Database for the functional annotation of proteins. *Nucleic Acids Res.*, **39**, D225–D229.
25. Gough,J., Karplus,K., Hughey,R. and Chothia,C. (2001) Assignment of homology to genome sequences using a library of hidden Markov models that represent all proteins of known structure. *J. Mol. Biol.*, **313**, 903–919.

the FEBS Journal

Effect of Ca^{2+} on the microtubule-severing enzyme p60-katanin

Insight into the substrate-dependent activation mechanism

Naoko Iwaya[1,2], Kohei Akiyama[2,3], Natsuko Goda[2,3,4], Takeshi Tenno[1,2,4], Yoshie Fujiwara[2,5,6], Daizo Hamada[2,5], Teikichi Ikura[7], Masahiro Shirakawa[1] and Hidekazu Hiroaki[2,3,4,5]

1 Department of Molecular Engineering, Graduate School of Engineering, Kyoto University, Japan
2 Division of Structural Biology, Graduate School of Medicine, Kobe University, Japan
3 Field of Supramolecular Biology, International Graduate School of Arts and Sciences, Yokohama City University, Japan
4 Structural Biology Research Center and Division of Biological Science, Nagoya University, Japan
5 Global Center of Excellence Program for Integrative Membrane Biology, Kobe University, Japan
6 Laboratory of Supramolecular Crystallography, Institute for Protein Research, Osaka University, Japan
7 Graduate School of Biomedical Science, Tokyo Medical and Dental University, Japan

Keywords
AAA ATPase; heteronuclear NMR; microtubule-severing enzyme; molecular modeling; three-dimensional domain rearrangement

Correspondence
H. Hidekazu, Structural Biology Research Center and Division of Biological Science, Graduate School of Science, Nagoya University, Furo-cho, Chikusa-ku, Nagoya, Aichi 464-8601, Japan
Fax: +81 52 789 4535
Tel: +81 52 747 6471
E-mail: hiroaki.hidekazu@f.mbox.nagoya-u.ac.jp

(Received 23 August 2011, revised 22 December 2011, accepted 8 February 2012)

doi:10.1111/j.1742-4658.2012.08528.x

Katanin p60 (p60-katanin) is a microtubule (MT)-severing enzyme and its activity is regulated by the p80 subunit (adaptor-p80). p60-katanin consists of an N-terminal domain, followed by a single ATPase associated with various cellular activities (AAA) domain. We have previously shown that the N-terminal domain serves as the binding site for MT, the substrate of p60-katanin. In this study, we show that the same domain shares another interface with the C-terminal domain of adaptor-p80. We further show that Ca^{2+} ions inhibit the MT-severing activity of p60-katanin, whereas the MT-binding activity is preserved in the presence of Ca^{2+}. In detail, the basal ATPase activity of p60-katanin is stimulated twofold by both MTs and the C-terminal domain of adaptor-p80, whereas Ca^{2+} reduces elevated ATPase activity to the basal level. We identify the Ca^{2+}-binding site at the end of helix 2 of the N-terminal domain, which is different from the MT-binding interface. On the basis of these observations, we propose a speculative model in which spatial rearrangement of the N-terminal domain relative to the C-terminal AAA domain may be important for productive ATP hydrolysis towards MT-severing. Our model can explain how Ca^{2+} regulates both severing and ATP hydrolysis activity, because the Ca^{2+}-binding site on the N-terminal domain moves close to the AAA domain during MT severing.

Structured digital abstract
- p80-CTD binds to p60-vMIT by pull down (View interaction)
- p80-CTD and p60-vMIT bind by nuclear magnetic resonance (View interaction)

Introduction

Katanin, spastin and fidgetin are three major microtubule (MT)-severing enzymes [1–5]. Katanin is a heterodimeric ATPase composed of two subunits, p60 (p60-katanin) and p80 (adaptor-p80). p60-katanin is the

Abbreviations
AAA, ATPase associated with various cellular activities; GST, glutathione *S*-transferase; MBP, maltose binding protein; MIT, microtubule interacting and trafficking; MT, microtubule; p60-vMIT, variant MIT domain of p60-katanin; p80-CTD, C-terminal domain of adaptor-p80.

catalytic factor responsible for breaking MTs in an ATP-dependent manner and contains a single ATPase associated with various cellular activities (AAA) domain. Adaptor-p80 is a regulatory factor containing six WD40 repeats, and is not essential for severing activity (Fig. 1) [6,7]. The MT-severing function of katanin is specific for certan cell types and cell cycles. For example, katanin promotes mitosis in proliferating cells, where it severs MTs at the mitotic spindle poles and increases the number of minus ends, resulting in the accumulation of γ-tubulin at the mitotic centrosomes [8]. However, katanin exhibits severing of the metaphase spindles at the spindle poles [9].

Recently, MT severing by katanin in neuronal cells has received increased attention, because of its contribution to neurite outgrowth and its implication in neurodegenerative diseases. In nascent neurons, it seems critical that the MTs are chopped by katanin at the cell body prior to being translocated into axons; subsequently, the chopped MT fragments disassemble into tubulin dimers that then reassemble into long MTs [10]. By contrast, excessive MT severing by exogenous overexpression of p60-katanin results in a shortening of the total process length of MTs in the hippocampal neuron [11]. In a model of neurodegenerative disease, MT breakdown by katanin triggers the loss of neurite spikes, but not the apoptosis of neurons, while the microtubule-associated protein, tau may protect MTs against attack by katanin [12]. Thus, the mechanism of p60-katanin regulation by adaptor-p80 or other factors such as tau in various cells and tissues, and at different developmental stages, is an important and unresolved issue.

We have previously shown that the substrate recognition mechanism is evolutionarily conserved between p60-katanin and its related AAA ATPase Vps4 [13,14]. Vps4 is responsible for the dissociation of membrane skeleton ESCRT-III fibrils. Note that p60-katanin and Vps4 share a common domain organization typical of type I AAA-ATPases, which consists of an N-terminal substrate-binding region, followed by a single AAA domain at the C-terminus. Despite the low (~ 20%) amino acid identity, we have shown that the 3D structure of the N-terminal domain of p60-katanin is strikingly similar to that of the microtubule interacting and trafficking (MIT) domain of Vps4 and other MIT

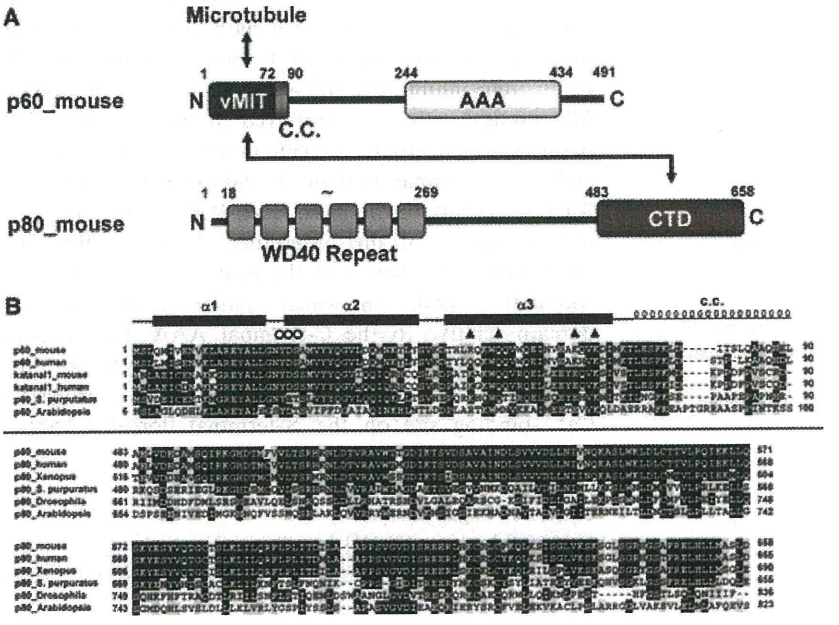

Fig. 1. Domain architectures and multiple sequence alignment of p60-katanin and adaptor-p80. (A) Domain architectures of mouse p60-katanin and adaptor-p80. The regions of interaction between each domain are indicated by an arrow. vMIT, variant MIT domain; cc, coiled-coil; AAA, AAA domain; CTD, C-terminal domain. (B) Multiple sequence alignment of p60-vMIT and p80-CTD, which interact with each other. The secondary structure elements of p60-vMIT are shown at the top. (▲) Residues involved in tubulin binding. (○) Residues involved in Ca^{2+} binding (see Fig. 6). The protein names and UniProtKB accession numbers are as follows: p60 mouse (Q9WV86), p60 human (O75449), katanal1 mouse (Q8K0T4), katanal1 human (Q9BW62), p60 *Strogylocentrotus purpuratus* (O61577), p60 *Arabidopsis* (Q9SEX2), p80 mouse (Q8BG40), p80 human (Q9BVA0), p80 *Xenopus* (Q4V7Y7), p80 *Strogylocentrotus purpuratus* (O61585), p80 *Drosophila* (Q9NHF0) and p80 *Arabidopsis* (Q8H0T9). The sequence alignment was generated using CLUSTAL X [45].

domains. Therefore, we named this domain 'variant MIT' (vMIT). Helices 2 and 3 of p60-vMIT make up the interface with MTs, which is similar that Vps4-MIT makes with its substrate ESCRT-III [14]. We further predicted that the binding site for p60-katanin in MTs is located at helix 12 of the α-tubulin subunit. Interestingly, this helix was recently proposed as the binding site for tau [15]. On the basis of our structural model of the p60–MT complex, competition between tau and katanin may occur at this 12th helix of α-tubulin. This unexpected coincidence is, however, consistent with the hypothesis proposed by Qiang *et al.* [12], in which the physiological role of tau is to protect the MT bundle in axons against the MT-severing activity of p60-katanin. Sudo & Baas [16] recently proposed this competition between tau (MT-stabilizing) and katanin (MT-severing) as a new therapeutic target for dementia and other neurodegenerative diseases.

In this study, we focused in detail on the role of the vMIT domain in regulating the ATPase activity of p60-katanin. It is known that either MT (substrate) alone or MT + p80 (adaptor) enhances the ATPase activity of p60-katanin [6,17,18], although the underlying molecular mechanism remains unclear. We therefore hypothesized that substrate/adaptor-dependent ATPase activation originates from p60-vMIT via a change in the location of this domain relative to the AAA domain. During the course of biochemical studies, we found that Ca^{2+} ions inhibit the MT-severing activity of p60-katanin. Furthermore, we found that Ca^{2+} directly affects the vMIT domain and then cancels the elevated ATPase activity of p60-katanin that occurs in the presence of either MT or the C-terminal domain of adaptor-p80 (p80-CTD). Given that Ca^{2+} is a key intracellular signal that indirectly regulates MT stability, as well as rearranging the cytoskeleton *in vivo*, we discuss the physiological implications of this finding.

Results

p80-CTD interacts with helix 1 of p60-vMIT

Previous biochemical studies by McNally and co-workers [7,19] showed that the N-terminal half of p60-katanin interacts with the C-terminal half of adaptor-p80, and in a previous study we showed from the 3D structure of p60-vMIT that the MT interface of p60-vMIT is located on the surface of helix 2/3. Therefore, we first examined the amino acid sequence of adaptor-p80 to identify the boundary of its C-terminal p60-binding domain. Figure 1 summarizes the domain organization of p60-katanin and adaptor-p80, and presents multiple sequence alignments of the two domains (p60-vMIT and p80-CTD) that interact with each other. Both domains are genetically conserved among many higher eukaryotes. Finally, we chose residues 480–655 of human p80 and 483–658 of mouse p80 for expression and purification because this region is markedly conserved among vertebrates (Fig. 1B).

Next, we determined the interface between p60-vMIT and p80-CTD using ^{1}H-^{15}N HSQC experiments. Binding of p80-CTD induced drastic line broadening of NH signals from ^{15}N-labeled p60-vMIT (Figs 2A and S1A). Mapping of the broadened signals upon binding to p80-CTD suggested that a wide area covering the surface of helices 1/2 and 1/3 of p60-vMIT might form the binding site for p80-CTD (Figs 2B and S1B; Table S1). We further examined the key residues involved in p80 binding by using Ala-substituted mutants of p60-vMIT in an *in vitro* pull-down assay (Fig. 2C). Wild-type p60-vMIT bound to p80-CTD in a ratio of almost 1 to 1, as judged from the density of the bands in SDS/PAGE (Fig. 2C, lanes 3 and 14). All Ala mutants also showed similar binding to p80-CTD. Therefore, we assumed that a polar single mutation does not completely inhibit the interaction between p60-vMIT and p80-CTD, probably because the interface is multivalent and covers a wide area, including helix 1.

Binding of either MT or p80-CTD to p60-vMIT stimulates the ATPase activity of p60-katanin

The ATPase activity of sea urchin and plant p60-katanin is stimulated by MTs alone, or MTs and adaptor-p80 [6,17,18]. In addition, the MT-severing activity of p60-katanin is stimulated by p80/con80 (412–655), according to a 4′-6-diamidino-2-phenylindole assay [6,7,18,20]. We performed experiments using mammalian full-length p60-katanin and p80-CTD, the domain that we isolated above. We investigated the ability of p60-katanin to hydrolyze ATP in the presence of either MTs or p80-CTD *in vitro* (Fig. 3). Under both conditions, the basal ATPase activity of p60-katanin was stimulated by MTs or p80-CTD in a concentration-dependent manner. In the presence of MTs, stimulation of the ATPase activity of p60-katanin reached almost a maximum at 0.5 μM MTs, corresponding to a p60/MT molar ratio of 0.8. We also found that the ATPase activity of p60-katanin is stimulated by the addition of p80-CTD alone. Stimulation of the ATPase activity of p60-katanin reached almost a maximum at 0.4 μM p80-CTD, corresponding to a p60/p80-CTD molar ratio of 1.0. Similar stimulation of p60-katanin was observed in the presence of both MTs and p80-CTD (Fig. S2A). These results show that binding

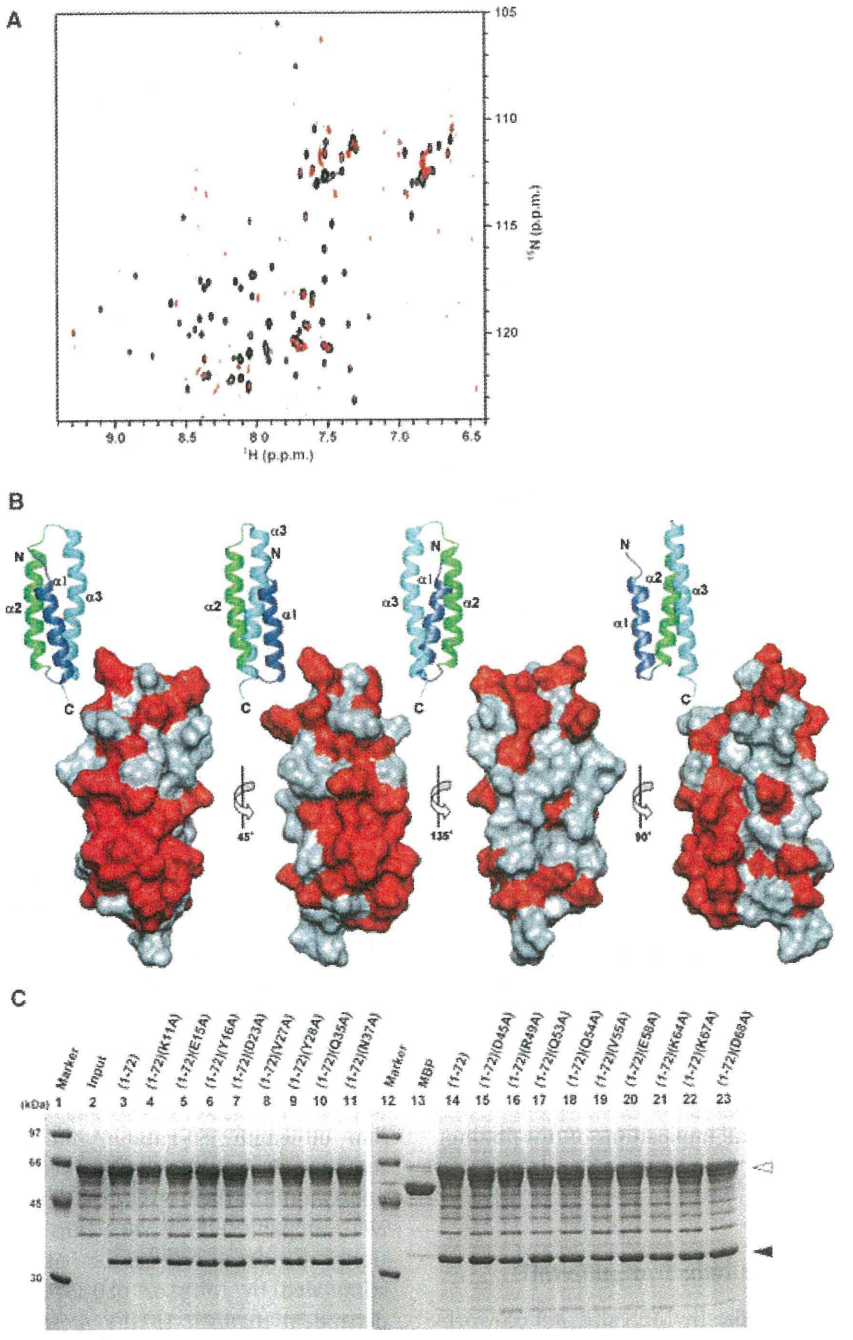

Fig. 2. Analyses of the interaction between p60-vMIT and p80-CTD. (A) ^{1}H-^{15}N HSQC spectra of p60-vMIT in the absence (black) and presence (red) of p80-CTD. Because a ^{1}H-^{15}N HSQC spectrum of p60-vMIT in the presence of p80-CTD was drastically broadening, it was represented at a high threshold level corresponding to the noise. (B) Residues whose NH signals of p60-vMIT were broadening in the presence of p80-CTD are mapped on the surface in red. The surface orientations are shown by the ribbon diagram of p60-vMIT (PDB: 2rpa). The residues in helix 1 had mostly been eliminated from the ^{1}H-^{15}N HSQC spectra, because those regions are involved in binding to p80-CTD. (C) Pull-down assays of MBP-tagged p80-CTD with GST-tagged p60-vMIT (wild-type and Ala mutants) *in vitro*. MBP-tagged p80-CTD was used as the input. Molecular sizes are shown in lanes 1 and 12. MBP was used as a negative control (lane 13). Recombinant proteins used for pull-down are indicated at the top of the gel. Filled and open arrowheads show MBP-tagged p80-CTD and GST-tagged p60-vMITs, respectively. The SDS/PAGE gel was stained with Coomassie blue.

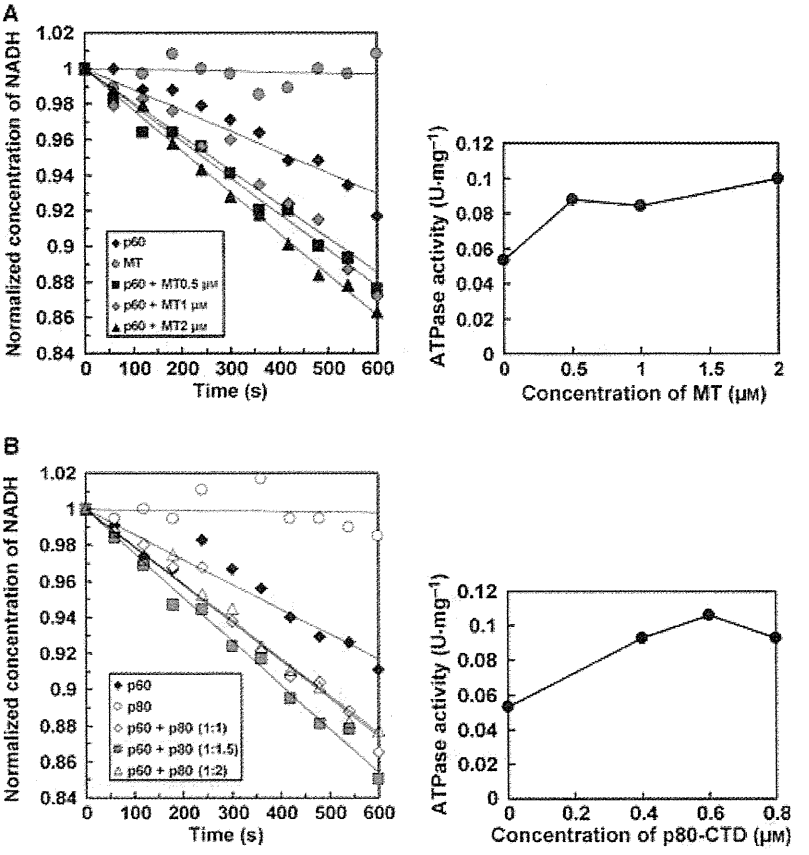

Fig. 3. The ATPase activity of full-length p60-katanin is stimulated in the presence of MT or p80-CTD. (A) ATPase activity of p60-katanin (0.4 μM) was monitored at 340 nm in the absence (black diamond) or presence of 0.5 μM (black square), 1 μM (gray diamond) and 2 μM (black triangle) of taxol-stabilized MTs (left). The gray circle indicates taxol-stabilized MTs (1 μM) without p60-katanin. ATPase activity was further expressed as μmoles of ATP hydrolyzed per min per mg of p60-katanin (units·mg^{-1}) in 0, 0.5, 1 and 2 μM of taxol-stabilized MTs (right). (B) ATPase activity of p60-katanin (0.4 μM) was monitored at 340 nm in the absence (black diamond) or presence of 0.4 μM (open diamond), 0.6 μM (gray square), and 0.8 μM (open triangle) of p80-CTD (left). Open circle indicates p80-CTD (0.4 μM) without p60-katanin. The ATPase activity was further expressed as micromoles of ATP hydrolyzed per min per mg of p60-katanin (units·mg^{-1}) in 0, 0.4, 0.6 and 0.8 μM of p80-CTD (right).

of either MTs or p80-CTD to the N-terminal domain of p60-katanin stimulates the basal ATPase activity of p60-katanin. In other words, p60-vMIT serves as a sensory regulator for the ATPase activity of p60-AAA.

Ca^{2+} negatively regulates the ATPase activity enhanced by MT or p80-CTD

Next, we examined the effect of Ca^{2+} on the ATPase activity of p60-katanin. Interestingly, we found that Ca^{2+} reduced the increase in ATPase activity induced by MTs or p80-CTD to the basal level (Fig. 4). Basal ATPase activity of p60-katanin in the absence of MTs or p80-CTD was not affected by Ca^{2+} (Fig. 4A), even at a higher concentration such as 10 mM (data not shown). In the presence of MTs, however, the ATPase activity of p60-katanin decreased as the Ca^{2+} concentration increased, finally decreasing to the basal level at 5 mM Ca^{2+} (Fig. 4B). A similar tendency was observed in the presence of p80-CTD, where the ATPase activity decreased to the basal level at 2 mM Ca^{2+} (Fig. 4C). These results suggest that Ca^{2+} regulates a putative On/Off mechanism for the ATPase activity of p60-katanin, which may be associated with either the vMIT or AAA domain.

Ca^{2+} inhibits the MT-severing activity of p60-katanin, but does not inhibit its binding to MTs

The dynamics of the polymerization–depolymerization of MTs is called 'dynamic instability' [21]. Growing

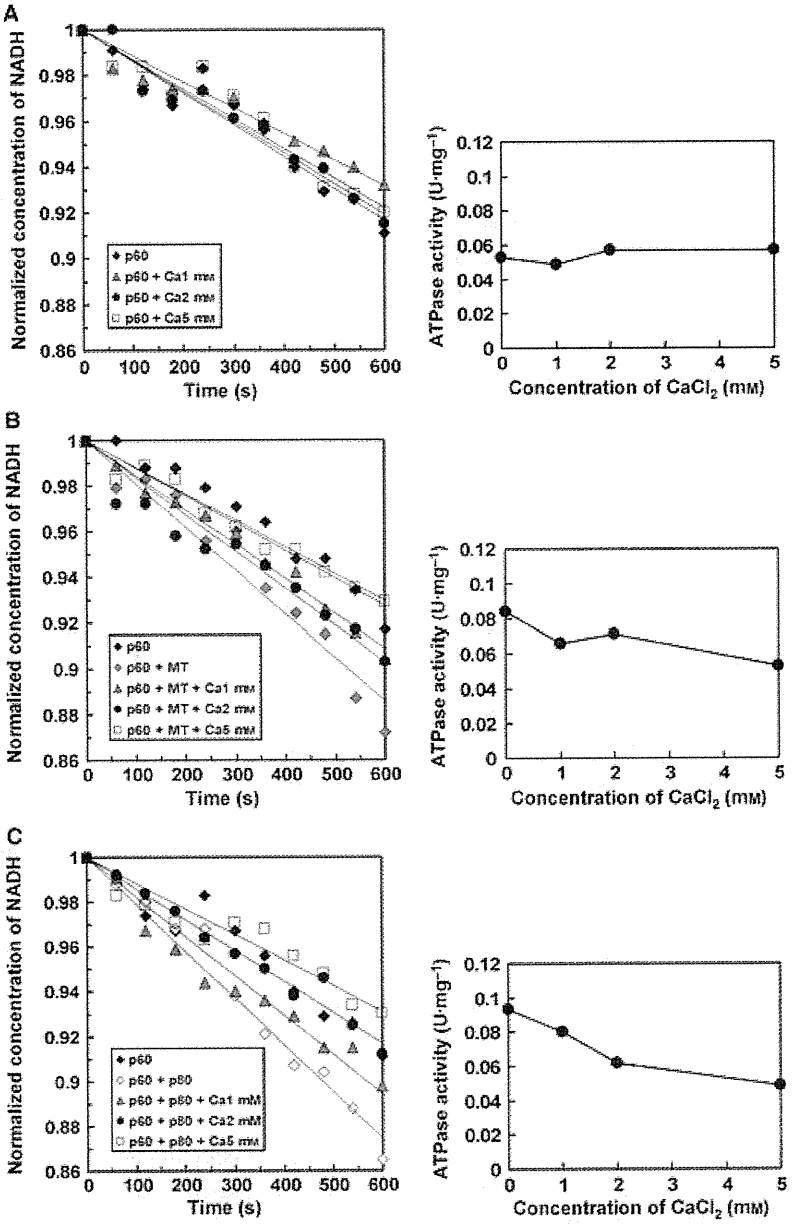

Fig. 4. Substrate-/adapter-stimulated ATPase activity of full-length p60-katanin is suppressed in the presence of Ca^{2+}. (A) ATPase activity of p60-katanin (0.4 μM) (black diamond) was monitored at 340 nm with increasing Ca^{2+} concentrations: 1 mM (gray triangle), 2 mM (black circle) and 5 mM (open square) (left). ATPase activity was further expressed as μmoles of ATP hydrolyzed per min per mg of p60-katanin (units·mg^{-1}) in 0, 1, 2 and 5 mM of Ca^{2+} (right). (B, C) ATPase activity of p60-katanin (0.4 μM) + taxol-stabilized MTs (1 μM) (gray diamond) (B) or p60-katanin (0.4 μM) + p80-CTD (0.4 μM) (open diamond) (C) was monitored at 340 nm with increasing Ca^{2+} concentrations: 1 mM (gray triangle), 2 mM (black circle) and 5 mM (open square) (left). ATPase activity in presence of MTs or p80-CTD was further expressed as μmoles of ATP hydrolyzed per min per mg of p60-katanin (units·mg^{-1}) in 0, 1, 2 and 5 mM of Ca^{2+} (right).

MTs suddenly start depolymerizing rapidly, known as a 'catastrophe'. However, shortening MTs often are 'rescued' and start polymerizing again [22]. *In vitro*, higher concentrations of Ca^{2+} in the mM range induce destabilization of MTs [23,24]; therefore, Ca^{2+} has been used for biochemical experiments to inhibit polymerization or induce the depolymerization of MTs. Figure 5A shows the effects of Ca^{2+} on the interaction