the FEBS Journal

Structure of the putative 32 kDa myrosinase-binding protein from *Arabidopsis* (At3g16450.1) determined by SAIL-NMR

Mitsuhiro Takeda[1], Nozomi Sugimori[2], Takuya Torizawa[2], Tsutomu Terauchi[2], Akira M. Ono[2], Hirokazu Yagi[3], Yoshiki Yamaguchi[3], Koichi Kato[3,4], Teppei Ikeya[2,5], JunGoo Jee[2], Peter Güntert[2,5,6], David J. Aceti[7], John L. Markley[7] and Masatsune Kainosho[1,2,5]

1 Graduate School of Science, Nagoya University, Japan
2 Graduate School of Science, Tokyo Metropolitan University, Hachioji, Japan
3 Graduate School of Pharmaceutical Sciences, Nagoya City University, Japan
4 Institute for Molecular Science, National Institute of Natural Sciences, Okazaki, Japan
5 Institute of Biophysical Chemistry and Center of Biomolecular Magnetic Resonance, Goethe University, Frankfurt am Main, Germany
6 Frankfurt Institute for Advanced Studies, Frankfurt am Main, Germany
7 Center for Eukaryotic Structural Genomics, Department of Biochemistry, University of Wisconsin-Madison, WI, USA

Keywords
lectin; myrosinase-binding protein; NMR structure; stereo-array isotope labeling; structural genomics

Correspondence
M. Kainosho, Graduate School of Science, Institute for Advanced Research, Furo-cho, Chikusa-ku, Nagoya 464-8601, Japan
Fax: +81 52 747 6433
Tel: +81 52 747 6474
E-mail: kainosho@nagoya-u.jp
J. L. Markley, Center for Eukaryotic Structural Genomics, Department of Biochemistry, University of Wisconsin-Madison, 433 Babcock Drive, Madison, WI 53706 1344, USA
Fax: +1 608 262 3759
Tel: +1 608 263 9349
E-mail: markley@nmrfam.wisc.edu

(Received 4 September 2008, revised 25 September 2008, accepted 29 September 2008)

doi:10.1111/j.1742-4658.2008.06717.x

The product of gene At3g16450.1 from *Arabidopsis thaliana* is a 32 kDa, 299-residue protein classified as resembling a myrosinase-binding protein (MyroBP). MyroBPs are found in plants as part of a complex with the glucosinolate-degrading enzyme myrosinase, and are suspected to play a role in myrosinase-dependent defense against pathogens. Many MyroBPs and MyroBP-related proteins are composed of repeated homologous sequences with unknown structure. We report here the three-dimensional structure of the At3g16450.1 protein from *Arabidopsis*, which consists of two tandem repeats. Because the size of the protein is larger than that amenable to high-throughput analysis by uniform $^{13}C/^{15}N$ labeling methods, we used stereo-array isotope labeling (SAIL) technology to prepare an optimally $^{2}H/^{13}C/^{15}N$-labeled sample. NMR data sets collected using the SAIL protein enabled us to assign ^{1}H, ^{13}C and ^{15}N chemical shifts to 95.5% of all atoms, even at a low concentration (0.2 mM) of protein product. We collected additional NOESY data and determined the three-dimensional structure using the CYANA software package. The structure, the first for a MyroBP family member, revealed that the At3g16450.1 protein consists of two independent but similar lectin-fold domains, each composed of three β-sheets.

The flowering plant *Arabidopsis thaliana* is an important model system for identifying plant genes and determining their functions. Analysis of the completed *Arabidopsis thaliana* genome revealed the presence of 25 498 genes encoding proteins from 11 000 families, including many new protein families [1]. To investigate the biological importance of these proteins, the Center for Eukaryotic Structural Genomics (CESG) at the University of Madison-Wisconsin has established platforms for protein structure determination by X-ray

Abbreviations
FAC, frontal affinity chromatography; MyroBP, myrosinase-binding protein; PA, pyridylamine; SAIL, stereo-array isotope labeling; UL, uniformly $^{13}C/^{15}N$-labeled.

crystallography and NMR spectroscopy, with protein production both by conventional heterologous gene expression in *Escherichia coli* and automated cell-free technology [2]. To date, targets for NMR analysis have been limited to proteins < 25 kDa, because this is the conventional size limit for high-throughput structure determination by NMR spectroscopy [2].

One of the motivations at CESG for choosing to develop a cell-free protein production platform was to be able to take advantage of the emerging new technology of optimal isotopic labeling for protein NMR spectroscopy. This approach, named stereo-array isotope labeling (SAIL), utilizes the incorporation of amino acids labeled with ^{2}H, ^{13}C and ^{15}N in order to minimize spectral complexity and spin diffusion within the protein while allowing detection of all connectivities required for sequence-specific assignments and determination of sufficient constraints for high-resolution solution structures [3]. The SAIL approach requires cell-free incorporation of the amino acids because the labeling patterns in the amino acids would become scrambled if they were incorporated in a cellular system [3]. As its first target for investigation by the SAIL approach, CESG chose the *A. thaliana* gene At3g16450.1, which encodes a 32 kDa, 299-residue protein with unknown structure.

At3g16450.1 has been classified as a myrosinase-binding protein-like protein. Myrosinase is a glucosinolate-degrading enzyme [4], and myrosinase-binding protein (MyroBP) has been identified as a component of high-molecular-mass myrosinase complexes in extracts of *Brassica napus* seed [5]. The presence of three myrosinase genes and several putative MyroBPs has been reported in *A. thaliana* [6–8]. The myrosinase/glucosinolate system is involved in plant defense against insects and pathogens [4], and hence MyroBP is implicated in this defense system, although experimental data supporting this notion are lacking [9]. Many MyroBPs and MyroBP-related proteins have a repetitive structure with two or more homologous sequences [10,11]. The homologous domains also have sequence similarity to some plant lectins, and, because seed MyroBP from *B. napus* has been found to bind to *p*-aminophenyl-α-D-mannopyranoside and to some extent to *N*-acetylglucosamine, the protein has been reported to possess lectin activity [10]. However, despite its functional importance, no three-dimensional structure has been determined for any domain of the MyroBP family.

We report here the three-dimensional structure of the At3g16450.1 protein, which consists of two homologous MyroBP-type domains. The structure, which was determined by NMR spectroscopy from a relatively low quantity of SAIL protein (approximately 60 nmol; 300 μL of 0.2 mM protein), revealed that At3g16450.1 consists of tandem lectin-like domains corresponding to the two homologous sequences (residues 1–144 and 153–299). To explore the sugar-binding activity of At3g16450.1, we investigated interactions between immobilized At3g16450.1 protein and fluorescently labeled (pyridylaminated, PA) sugars by frontal affinity chromatography (FAC) [12]. Of the carbohydrates tested, only a few PA sugars showed significant affinity for the immobilized At3g16450.1. This result is discussed in light of the possible biological function of this protein. This study demonstrates the power of the SAIL approach in determining the structure of a larger protein by semi-automated means and with a minimal amount of material. It also shows how a structure determined by NMR spectroscopy can be the springboard for easily performed functional investigations.

Results

Preparation of SAIL At3g16450.1

At3g16450.1 is a 299-residue protein with a molecular weight of 32 kDa. In our earlier work [13], we assigned the backbone resonances of At3g16450.1 using samples labeled uniformly with ^{13}C/^{15}N or ^{2}H/^{13}C/^{15}N. However, further progress towards structure determination was impeded by the problems of spectral crowding and broadened signals, as commonly seen in the NMR spectra of uniformly ^{13}C/^{15}N-labeled (UL) large proteins. In the present study, we used the SAIL technique [3] to address these problems. As an initial step, we optimized the conditions for *E. coli* cell-free production of At3g16450.1 with regard to reaction temperature, duration of incubation, and expression vector. For comparison purposes, [U-^{13}C,U-^{15}N]-labeled At3g16450.1 (UL At3g16450.1) was prepared using an *E. coli in vivo* expression system.

Fig. 1. Comparison of ^{1}H-^{13}C constant-time HSQC NMR spectra of 0.6 mM of UL At3g16450.1 and 0.2 mM of SAIL At3g16450.1. (A) Full spectrum of UL At3g16450.1. (B) Full spectrum of SAIL At3g16450.1. (C) Methylene region of UL At3g16450.1. (D) Methylene region of SAIL At3g16450.1. (E) Methyl region of UL At3g16450.1. (F) Methyl region of SAIL At3g16450.1. Spectra were recorded at 27.5°C at ^{1}H frequency of 800 MHz. In the case of the SAIL protein, ^{2}H decoupling was applied during the ^{13}C chemical shift evolution.

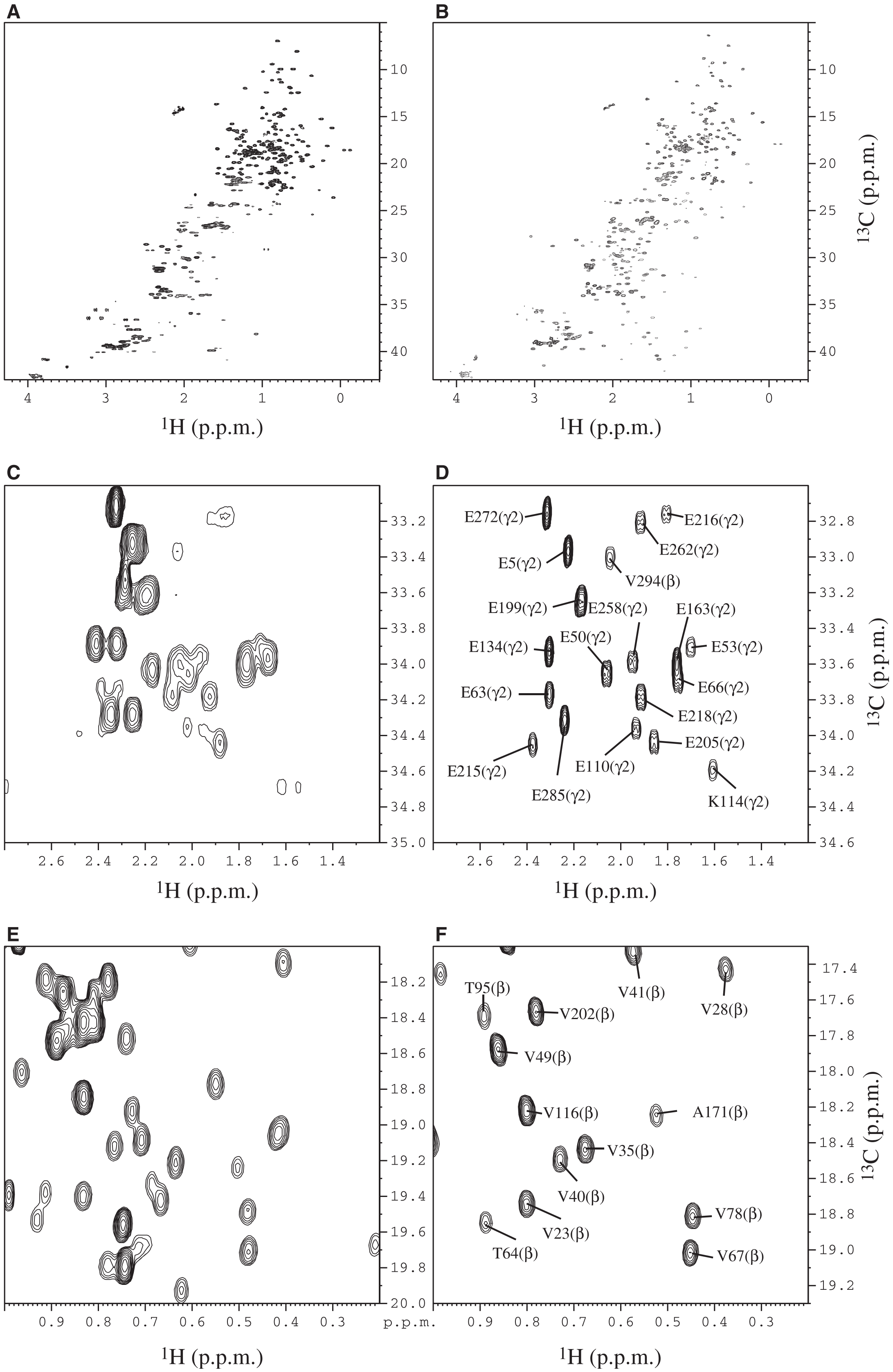

A
B
C
D
E
F
^{1}H (p.p.m.)
^{13}C (p.p.m.)
p.p.m.
E272(γ2)
E216(γ2)
E262(γ2)
E5(γ2)
V294(β)
E199(γ2)
E258(γ2)
E163(γ2)
E50(γ2)
E134(γ2)
E53(γ2)
E66(γ2)
E63(γ2)
E218(γ2)
E205(γ2)
E110(γ2)
E215(γ2)
E285(γ2)
K114(γ2)
T95(β)
V41(β)
V28(β)
V202(β)
V49(β)
V116(β)
A171(β)
V35(β)
V40(β)
V78(β)
V23(β)
T64(β)
V67(β)

Table 1. NMR constraints and structure calculation statistics for At3g16450.1[a].

Completeness of the chemical shift assignments (%)	
All	95.5
Backbone	97.8
Side chain	93.3
NOE distance constraints	
Total	1982
Short-range, $\|i-j\| \leq 1$	1192
Medium-range, $1 < \|i-j\| < 5$	111
Long-range, $\|i-j\| \geq 5$, intra-molecular	679
Maximal violation (Å)	0.18
Torsion angle constraints	
ϕ	138
ψ	136
Maximal violation (°)	2.6
Restrained hydrogen bonds	124
CYANA target function value (Å^2)	1.77 ± 0.56
AMBER energies (kcal·mol^{-1})	
Total	−7508 ± 21
van der Waals	−2239 ± 30
Ramachandran plot statistics (%) [35]	
Residues in most favored regions	89.0
Residues in additional allowed regions	9.5
Residues in generously allowed regions	1.0
Residues in disallowed regions	0.5
Root mean square deviation from the averaged coordinates (Å)	
Backbone atoms of residues 2–144 (N-domain)	1.12 ± 0.19
Heavy atoms of residues 2–144 (N-domain)	1.65 ± 0.16
Backbone atoms of residues 153–297 (C-domain)	0.69 ± 0.10
Heavy atoms of residues 153–297 (C-domain)	1.08 ± 0.09

[a] The completeness of the ^{1}H, ^{13}C and ^{15}N chemical shift assignments was evaluated for the aliphatic, aromatic, backbone amide and Asn/Gln/Trp side-chain amide nuclei, excluding the carbon and nitrogen atoms not bound to ^{1}H. Where applicable, the value given corresponds to the average over the 20 energy-refined conformers that represent the solution structure. CYANA target function values were calculated before energy refinement.

Comparison of NMR spectra of SAIL and UL At3g16450.1

Although the concentration of the SAIL protein was lower than that of the UL protein by a factor of three (SAIL, 0.2 mM; UL, 0.6 mM), the NMR spectra of SAIL At3g16450.1 exhibited higher signal-to-noise ratios than those of UL At3g16450.1. The ^{1}H-^{13}C constant-time HSQC spectrum of SAIL At3g16450.1 was less crowded and better resolved than that of UL At3g16450.1 (Fig. 1A,B). The extensive stereo- and regio-specific deuteration of the SAIL protein led to diminished overlaps and sharpened peaks, particularly in the methylene region, without compromising essential structural information (Fig. 1C,D). In the methyl region, the regio-specifically labeled methyl resonances from the SAIL sample were much less crowded (Fig. 1E,F). As a result of these striking spectral improvements, it became possible to use established methods [14] to assign 95.5% of the resonances of SAIL At3g16450.1. The chemical shifts for SAIL At3g16450.1 have been deposited in the Biological Magnetic Resonance Data Bank (BMRB) [15] with accession number 15607. In addition, 93% of the backbone carbonyl ^{13}C shifts had been assigned previously using uniformly ^{13}C/^{15}N-labeled protein [13]. These assigned chemical shifts were used as input for the TALOS program [16] to obtain dihedral angle constraints.

Solution structure of SAIL At3g16450.1

Assignment of the NOE peaks of At3g16450.1 and the structure determination were accomplished by use of the CYANA program [17,18]. The structural statistics are summarized in Table 1. Although the 20 conformers representing the structures of At3g16450.1 did not superimpose well when the full sequence was considered (residues 1-299), each individual domain (residues 1-144 or residues 153-299) superimposed well when considered separately (Fig. 2A,B). Residues 16–21 and 45–47 exhibited severe line broadening, probably arising from internal dynamics of these residues on the intermediate time scale for chemical shifts. As a result, these are the least well-defined regions of the N-terminal domain. The C-terminal domain yielded reasonably well-converged structures, including the side-chain conformations of residues in its core (Fig. 2C,D).

Residues 145–152 in the linker region between the two domains are highly disordered. In addition, a careful search failed to reveal any inter-domain NOE peaks. Thus the relative orientations of the two domains appear not to be fixed, and the overall structure of At3g16450.1 is best described as two tandem structural domains connected by a flexible linker (Fig. 3A). The secondary structural elements of At3g16450.1, extracted from the coordinates of the three-dimensional structure using the DSSP algorithm [19], showed that each domain has a similar structure consisting of three β-sheets related by pseudo three-fold symmetry (Fig. 3B).

The coordinates of the 20 energy-refined conformers that represent the solution structure of At3g16450.1 have been deposited in the Protein Data Bank with accession code 2JZ4. A structural homology search using the program DALI at the European Molecular Biology Laboratory (EMBL) [20,21] yielded the agglutinin from *Maclura promifera* (Protein Data Bank code

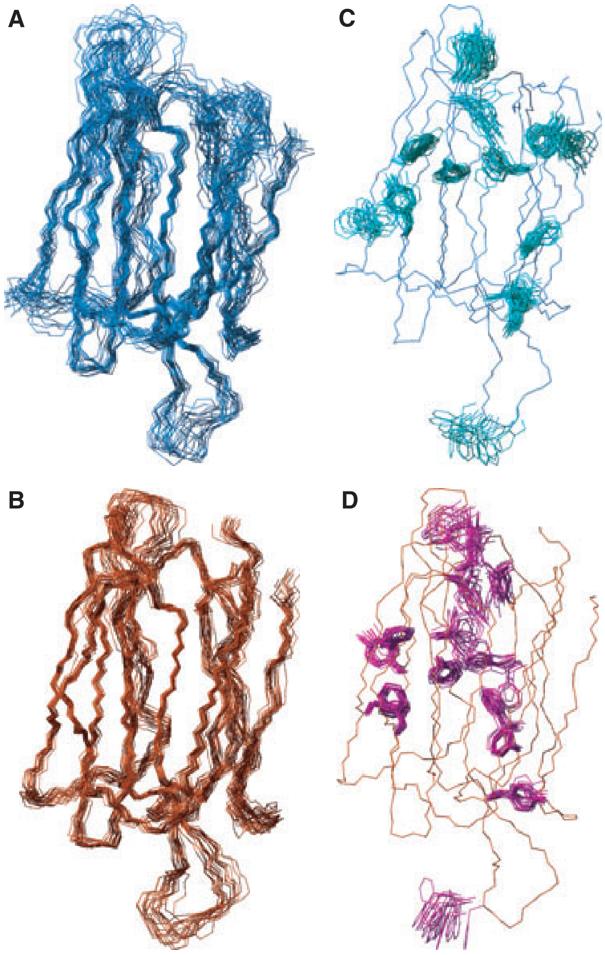


Fig. 2. Three-dimensional NMR structure of At3g16450.1. (A) Superposition of the 20 energy-minimized conformers that represent the 3D solution structure of the N-terminal domain. (B) Superposition of conformers representing the C-terminal domain. (C) Aromatic side chains and one backbone trace of the NMR structures for the N-terminal domain. (D) Aromatic side chains and one backbone trace of the NMR structure of the C-terminal domain.

1JOT), a plant lectin, as the closest structure. The root mean square deviation values for the N- and C-terminal domains versus the agglutinin are 2.2 and 2.0Å, respectively. Thus each of the two domains of At3g16450.1 adopts a lectin fold. The orientation of the N-terminal domain relative to the C-terminal domain could not be defined owing to the absence of inter-domain NOEs. To confirm the molecular organization of the tandem arrangement, expression vectors were constructed that separately encoded the N-terminal half (residues 1–153) and the C-terminal half (residues 151–299) of At3g16450.1, and these were used to prepare ^{15}N-labeled samples of each domain. The ^{1}H-^{15}N HSQC spectrum of each domain was well dispersed, and, when overlaid, closely approximated the spectrum of full-length At3g16450.1 (Fig. 4A,B). This result confirms the structural arrangement of At3g16450.1 as two independent tandem structural domains.

Interaction analysis of At3g16450.1 with sugars

Because each structural domain of At3g16450.1 was found to adopt a lectin fold, we assayed At3g16450.1 for possible sugar-binding activity. We utilized 13 fluorescence-labeled oligosaccharides (PA sugars) as candidates. Four PA sugars eluted more slowly than the tetra-sialyl PA-glycan as a control PA sugars from a column of immobilized At3g16450.1 (Fig. 5A,B and Table 2). On the basis of the elution profiles, the K_d values for the four PA sugars to At3g16450.1 were estimated to be low, at most 10^{-4} M. To further examine the observed interaction, we acquired ^{1}H-^{15}N HSQC spectra of ^{15}N-labeled At3g16450.1 in the presence and absence of maltohexaose, (Glcα1-4Glc)$_3$. However, addition of (Glcα1-4Glc)$_3$ did not cause any perturbation of NMR resonances, even when the concentration of the sugar was ten times higher than that of the protein (data not shown). By contrast, NMR titration of At3g16450.1 with (Glcα1-4Glc)$_3$-PA led to distinct chemical shift changes for certain NMR resonances (Fig. 5C), but addition of PA as the ligand resulted only in limited subtle changes. These results suggest that both PA and the (Glcα1-4Glc)$_3$ elements contribute to the observed interactions. Residues in both the N- and C-terminal domains of At3g16450.1 were affected by the presence of PA sugars (Fig. 5C, blue and red boxes). Taken together, these binding analyses suggest that At3g16450.1 has the potential to bind PA sugars with specificity for the sugar structure, although none of the various sugars tested exhibited a strong affinity.

Discussion

In this study, we determined the solution structure of the 32 kDa At3g16450.1 protein from *A. thaliana* by the SAIL-NMR method. This is the first application of SAIL-NMR in a structural genomics study. It provided the first structure for a member of the hitherto structurally unexplored MyroBP family.

At3g16450.1 consists of two tandem domains, each composed of three β-sheets. The fold of each domain is nearly identical to that of an agglutinin (Protein Data Bank code 1JOT), which shares sequence identities of 26 and 33% with the N- and C-terminal domains of At3g16450.1, respectively. Sequence similarity searches performed by PSI-BLAST [22] identified other MyroBPs and MyroBP-like proteins from *A. thaliana* and *B. napus*, with sequence identities to

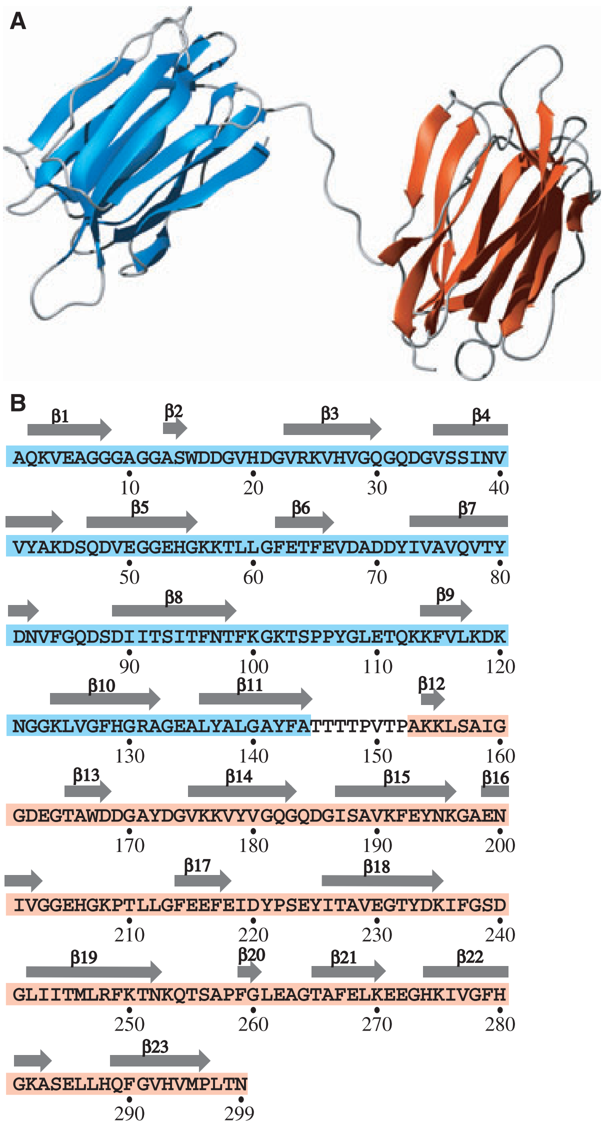

A
B
β1
β2
β3
β4
AQKVEAGGGAGGASWDDGVHDGVRKVHVGQGQDGVSSINV
10
20
30
40
β5
β6
β7
VYAKDSQDVEGGEHGKKTLLGFETFEVDADDYIVAVQVTY
50
60
70
80
β8
β9
DNVFGQDSDIITSITFNTFKGKTSPPYGLETQKKFVLKDK
90
100
110
120
β10
β11
β12
NGGKLVGFHGRAGEALYALGAYFATTTTPVTPAKKLSAIG
130
140
150
160
β13
β14
β15
β16
GDEGTAWDDGAYDGVKKVYVGQGQDGISAVKFEYNKGAEN
170
180
190
200
β17
β18
IVGGEHGKPTLLGFEEFEIDYPSEYITAVEGTYDKIFGSD
210
220
230
240
β19
β20
β21
β22
GLIITMLRFKTNKQTSAPFGLEAGTAFELKEEGHKIVGFH
250
260
270
280
β23
GKASELLHQFGVHVMPLTN
290
299

Fig. 3. Secondary structure of At3g16450.1. (A) Ribbon representation of the NMR structure of At3g16450.1. These figures were prepared using MOLMOL [25]. Due to the lack of NOEs, the relative orientation between the N- and C-terminal domains could not be defined. (B) Primary sequence of At3g16450.1. The sequences that correspond to the N-terminal (residues 1-144) and C-terminal (residues 153-299) structural domains are highlighted in blue and pink, respectively, and β-strands are indicated by arrows above the sequence.

the At3g16450.1 domains ranging from 30% to 70%. The most highly conserved regions correspond to the β-strands (Fig. 6). The N- and C-terminal domains of At3g16450.1, with 51% sequence identity to each other, are superimposed with root mean square deviations of 1.3 Å for the backbone of the β-strands and 1.7 Å if the loop regions are included, indicating that all of these family members adopt a similar fold.

It has been reported that seed MyroBP from *B. napus* possesses lectin activity, binding to *p*-aminophenyl-α-D-mannopyranoside and to some extent to *N*-acetylglucosamine [10]. Because myrosinase contains potential N-linked sugar-binding sites [23], the sugar-binding activity of MyroBP is implicated in binding to myrosinase. In the case of At3g16450.1, the protein did not show a significant affinity for sugar structures specific to N-linked glycan, but rather showed weak affinity for starch or glycolipid, raising the possibility that the lectin activity of the MyroBP family is also involved in interaction between a myrosinase complex and other molecules. It is also noteworthy that a UniGene database search [24] suggested that At3g16450.1 is expressed in leaf and root. Because myrosinases have also been shown to be expressed in *A. thaliana* leaf [6,8], it may be suspected that At3g16450.1 forms a complex with myrosinase, thereby guiding the myrosinase to a damaged site in the leaf via weak interactions with starch in the leaf or glycolipid from foreign pathogens. However, it is obvious that further study will be required to determine the biological importance of MyroBP–sugar interactions.

Many MyroBP and MyroBP-related proteins contain tandem lectin domains as shown in Fig. 6. The tandem domains present in MyroBP family members may participate in multivalent sugar binding as observed with other carbohydrate binding proteins with multiple domains. Results of the NMR chemical-shift perturbation experiments (Fig. 5C) suggest that both domains of At3g16450.1 can participate in a bivalent sugar binding. It is also probable that each homologous domain of the MyroBP family possesses different ligand-binding properties, thereby providing a broad binding specificity. In some proteins containing tandem homologous domains, inter-domain interactions fix the relative orientation of the domains in a specific multi-domain structure that is essential for biological function. Other proteins with tandem domains contain a flexible linker, and a specific structure may be adopted only when a target is bound. The present study suggests that At3g16450.1 belongs to the latter category.

The major problems with structural genomics studies using NMR are low solubility and molecular-weight limitations [2]. As shown by this study, the SAIL-NMR method provides a promising approach to overcoming both of these problems. One important aspect of the SAIL technology is that the signal intensities for the SAIL protein are several times stronger than for the corresponding UL sample [3], thus making it possible to perform structure determination for proteins even at low concentration. In this study, the structure was determined using a 0.2 mM sample of SAIL

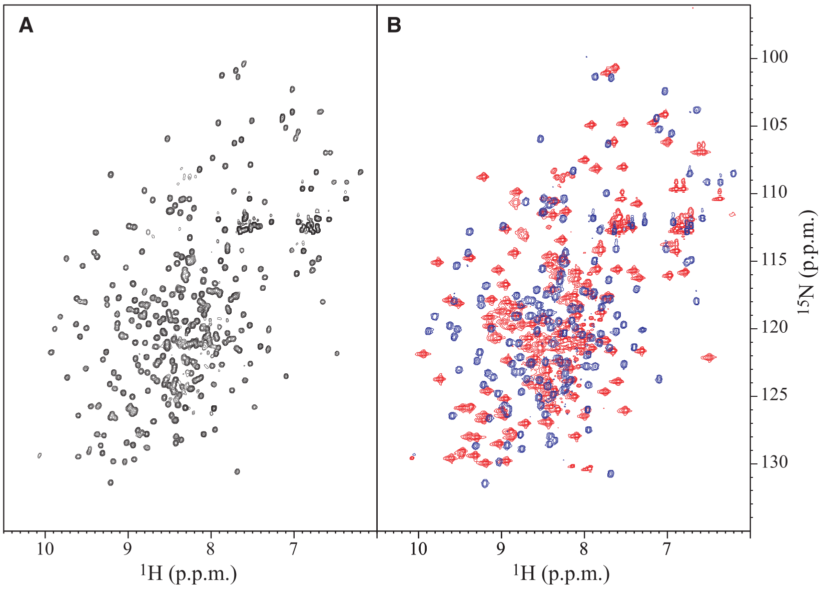


Fig. 4. Comparison of the NMR spectra of full-length At3g16450.1 and its isolated N- and C-terminal halves. (A) ^{1}H-^{15}N HSQC spectrum of full-length (residues 1–299) SAIL At3g16450.1. (B) Overlay of ^{1}H-^{15}N HSQC spectra of the N-terminal (residues 1–153, blue) and C-terminal (residues 151–299, red) halves of [U-^{15}N]-labeled At3g16450.1. These spectra were acquired at 27.5°C, pH 6.8, using a Bruker DRX600 NMR spectrometer. The pattern of the overlaid spectra is almost identical to that of the full-length construct, showing that the two domains of At3g16450.1 are largely independent.

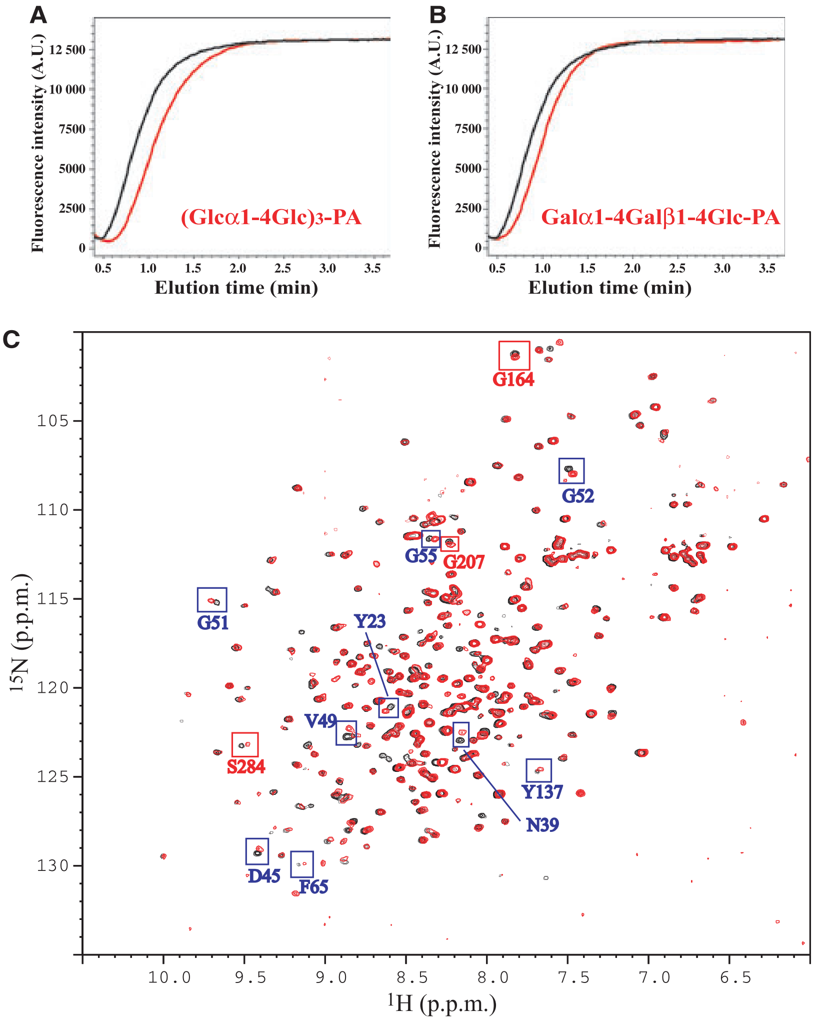


Fig. 5. Investigation of sugar-binding properties of At3g16450.1. (A) Elution profile from the FAC binding assay for (Glcα1-4Glc)$_3$-PA (red) and control sugar (black). (B) FAC binding assay for Galα1-4Galβ1-4Glc-PA (red) and control PA sugar (black). (C) Overlay of the ^{1}H-^{15}N HSQC spectra of uniformly ^{15}N-labeled At3g16450.1 in the absence (black) and presence (red) of (Glcα1-4Glc)$_3$-PA. Assignments and boxes (blue for the N-terminal domain; red for the C-terminal domain) indicate some of the perturbed resonances.

Table 2. Summary of results of the FAC binding assay for At3g16450.1 with various PA sugars.

	Major natural location
PA sugars that showed affinity for At3g16450.1	
(Glcα1-4Glc)$_3$ maltohexaose	Starch of higher plants
(Glcα1-6Glc)$_3$ isomaltohexaose	Starch of higher plants
Galα1-4Galβ1-4Glc	Glycolipid
GalNAcα1-3(Fucα1-2) Galβ1-3(Fucα1-4)GlcNAcβ1-3Galβ1-4Glc	Glycolipid
PA sugars that did not show affinity for At3g16450.1	
Galβ1-3(Fucα1-4)GlcNAcβ1-3Galβ1-4Glc	Glycolipid
Galβ1-4(Fucα1-3)GlcNAcβ1-3Galβ1-4Glc	Glycolipid
(GlcNAcβ1-4GlcNAc)$_3$ Chitohexaose	Insects and crustaceans
(Glcβ1-4Glc)$_3$ Cellohexaose	Cell walls of higher plants
(Glcβ1-3Glc)$_3$ Laminarihexaose	Pachyman of *Poria cocos*
Man9GN2 (high-mannose type) (code no. M9.1)	*N*-glycan
GlcNAcβ1-2Manα1-6 (GlcNAcβ1-2Manα1-3) Manβ1-4GlcNAcβ1-4(Fucα1-6) GlcNAc (code no. 210.1)	*N*-glycan
Galβ1-4GlcNAcβ1-2Mana1-6 (Galβ1-4GlcNAcβ1-2 Manα1-3)Manβ1-4GlcNAcβ1-4(Fucα1-6) GlcNAc (code no. 210.4)	*N*-glycan
GlcNAcβ1-2Manα1-6(GlcNAcβ1-2Manα1-3) Manβ1-4(Xylβ1-2)GlcNAcβ1-4 (Fucα1-3)GlcNAc (code no. 210.1FX)	*N*-glycan

At3g16450.1. The SAIL-NMR method offers the opportunity to determine structures of proteins with low solubility or poor yield. The SAIL method can also accelerate the process of structural analysis. The spectral simplification achieved by SAIL with this larger protein makes it possible to use semi- or fully automated methods developed for use with smaller proteins to analyze the NMR data. We are developing a software package that exploits the benefits of the SAIL method [25–27]. Finally, the SAIL method is expected to enable functional investigations of larger proteins.

Experimental procedures

Plasmid construction

The construction of pET15b (Novagen, Madison, WI, USA) harboring At3g16450.1 was performed as described previously [13]. The vector used for cell-free production of At3g16450.1 was constructed according to a strategy described previously [28]. DNA coding for the N-terminal histidine tag followed by the At3g16450.1 was subcloned into pIVEX2.3d (Roche, Pleasanton, CA, USA) between the *Nco*I/*Nde*I and *Nde*I/*Bam*HI sites, respectively. Silent mutations were introduced into the N-terminal sequence to enhance the expression rate [28]. Expression vectors coding for the N-terminal (residues 1–153) and C-terminal (residues 151–299) domains of At3g16450.1 were constructed by cloning the corresponding target sequence into the *Nde*I and *Bam*HI sites of pET15b.

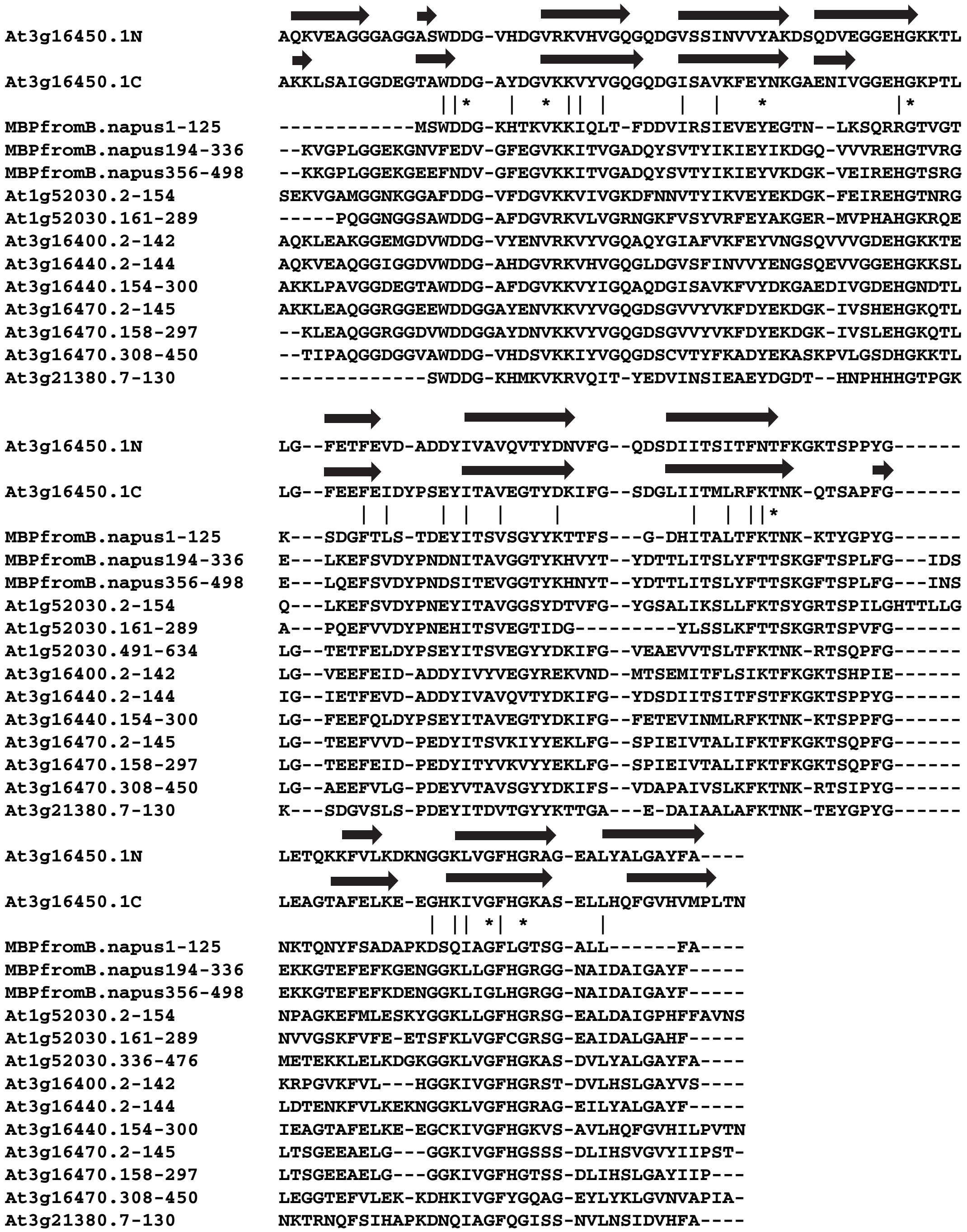
Preparation of labeled proteins

[U-^{15}N]- and [U-^{13}C, U-^{15}N]-labeled proteins were produced by culturing *Escherichia coli* BL21 (DE3) strain harboring the corresponding expression vector in M9 medium containing ^{15}NH$_4$Cl and/or [U-^{13}C]-labeled glucose as the sole nitrogen and carbon sources. Cells were cultured at 30 °C with shaking. Expression was induced by the addition of isopropyl thio-β-D-galactoside (IPTG) at a final concentration of 1 mM, and cells were harvested 6.5 h after induction.

SAIL At3g16450.1 was produced by *E. coli* cell-free expression. A total of 110 mg of SAIL amino acid mixture was used, with the amount of each individual SAIL amino acid proportional to the amino acid composition of At3g16450.1. A home-made *E. coli* S30 extract was used, and the reaction was performed as previously described [25,28]. The volumes of the inner and outer solutions were 10 and 40 mL, respectively. The reaction was carried out at 30 °C for 15 h with shaking. To prevent degradation of the produced protein, a protease inhibitor cocktail (Roche) was added to the reaction. The At3g16450.1 protein was purified as described previously [13].

NMR spectroscopy

The NMR sample used for the structure determination contained 0.2 mM SAIL At3g16450.1 protein in 20 mM bis-Tris(2-carboxymethyl)phosphine: HCl(D19, 98%) (Cambridge Isotope Laboratories Andover, MA, USA), 100 mM KCl, 10% D$_2$O, pH 6.8. NMR spectra were recorded on a Bruker (Tsukuba, Japan) Avance 600 MHz spectrometer equipped with a 5 mm ^{1}H-observe triple-resonance cryogenic probe (Bruker TXI cryoProbe), and on a Bruker Avance 800 MHz spectrometer at 27.5 °C. The spectra were processed using the programs XWINNMR version 3.5 (Bruker) or NMRPIPE [29], and analyzed using the program SPARKY (T. D. Goddard and D. G. Kneller, Department of Pharmaceutical Chemistry, University of California, San Francisco, CA, USA). Backbone and β-CH resonances were assigned using 2D HSQC, and 3D HN(CO)CACB and HBHA(CO)NH spectra. Side-chain resonances were assigned using 3D H(CCCO)NH, (H)CC(CO)NH, HCCH-TOCSY, constant time-HCCH-COSY, ^{13}C-edited NOESY



```
At3g16450.1N             AQKVEAGGGAGGASWDDG-VHDGVRKVHVGQGQDGVSSINVVYAKDSQDVEGGEHGKKTL
At3g16450.1C             AKKLSAIGGDEGTAWDDG-AYDGVKKVYVGQGQDGISAVKFEYNKGAENIVGGEHGKPTL
                                       ||*   |  * || |      |  |   *           |*
MBPfromB.napus1-125      ------------MSWDDG-KHTKVKKIQLT-FDDVIRSIEVEYEGTN--LKSQRRGTVGT
MBPfromB.napus194-336    --KVGPLGGEKGNVFEDV-GFEGVKKITVGADQYSVTYIKIEYIKDGQ-VVVREHGTVRG
MBPfromB.napus356-498    --KKGPLGGEKGEEFNDV-GFEGVKKITVGADQYSVTYIKIEYVKDGK-VEIREHGTSRG
At1g52030.2-154          SEKVGAMGGNKGGAFDDG-VFDGVKKVIVGKDFNNVTYIKVEYEKDGK-FEIREHGTNRG
At1g52030.161-289        -----PQGGNGGSAWDDG-AFDGVRKVLVGRNGKFVSYVRFEYAKGER-MVPHAHGKRQE
At3g16400.2-142          AQKLEAKGGEMGDVWDDG-VYENVRKVYVGQAQYGIAFVKFEYVNGSQVVVGDEHGKKTE
At3g16440.2-144          AQKVEAQGGIGGDVWDDG-AHDGVRKVHVGQGLDGVSFINVVYENGSQEVVGGEHGKKSL
At3g16440.154-300        AKKLPAVGGDEGTAWDDG-AFDGVKKVYIGQAQDGISAVKFVYDKGAEDIVGDEHGNDTL
At3g16470.2-145          AKKLEAQGGRGGEEWDDGGAYENVKKVYVGQGDSGVVYVKFDYEKDGK-IVSHEHGKQTL
At3g16470.158-297        --KLEAQGGRGGDVWDDGGAYDNVKKVYVGQGDSGVVYVKFDYEKDGK-IVSLEHGKQTL
At3g16470.308-450        --TIPAQGGDGGVAWDDG-VHDSVKKIYVGQGDSCVTYFKADYEKASKPVLGSDHGKKTL
At3g21380.7-130          -------------SWDDG-KHMKVKRVQIT-YEDVINSIEAEYDGDT--HNPHHHGTPGK

At3g16450.1N             LG--FETFEVD-ADDYIVAVQVTYDNVFG--QDSDIITSITFNTFKGKTSPPYG------
At3g16450.1C             LG--FEEFEIDYPSEYITAVEGTYDKIFG--SDGLIITMLRFKTNK-QTSAPFG------
                                | |    | |  |    |           |  | ||*
MBPfromB.napus1-125      K---SDGFTLS-TDEYITSVSGYYKTTFS---G-DHITALTFKTNK-KTYGPYG------
MBPfromB.napus194-336    E---LKEFSVDYPNDNITAVGGTYKHVYT--YDTTLITSLYFTTSKGFTSPLFG---IDS
MBPfromB.napus356-498    E---LQEFSVDYPNDSITEVGGTYKHNYT--YDTTLITSLYFTTSKGFTSPLFG---INS
At1g52030.2-154          Q---LKEFSVDYPNEYITAVGGSYDTVFG--YGSALIKSLLFKTSYGRTSPILGHTTLLG
At1g52030.161-289        A---PQEFVVDYPNEHITSVEGTIDG---------YLSSLKFTTSKGRTSPVFG------
At1g52030.491-634        LG--TETFELDYPSEYITSVEGYYDKIFG--VEAEVVTSLTFKTNK-RTSQPFG------
At3g16400.2-142          LG--VEEFEID-ADDYIVYVEGYREKVND--MTSEMITFLSIKTFKGKTSHPIE------
At3g16440.2-144          IG--IETFEVD-ADDYIVAVQVTYDKIFG--YDSDIITSITFSTFKGKTSPPYG------
At3g16440.154-300        LG--FEEFQLDYPSEYITAVEGTYDKIFG--FETEVINMLRFKTNK-KTSPPFG------
At3g16470.2-145          LG--TEEFVVD-PEDYITSVKIYYEKLFG--SPIEIVTALIFKTFKGKTSQPFG------
At3g16470.158-297        LG--TEEFEID-PEDYITYVKVYYEKLFG--SPIEIVTALIFKTFKGKTSQPFG------
At3g16470.308-450        LG--AEEFVLG-PDEYVTAVSGYYDKIFS--VDAPAIVSLKFKTNK-RTSIPYG------
At3g21380.7-130          K---SDGVSLS-PDEYITDVTGYYKTTGA---E-DAIAALAFKTNK-TEYGPYG------

At3g16450.1N             LETQKKFVLKDKNGGKLVGFHGRAG-EALYALGAYFA----
At3g16450.1C             LEAGTAFELKE-EGHKIVGFHGKAS-ELLHQFGVHVMPLTN
                                      | || *| *      |
MBPfromB.napus1-125      NKTQNYFSADAPKDSQIAGFLGTSG-ALL------FA----
MBPfromB.napus194-336    EKKGTEFEFKGENGGKLLGFHGRGG-NAIDAIGAYF-----
MBPfromB.napus356-498    EKKGTEFEFKDENGGKLIGLHGRGG-NAIDAIGAYF-----
At1g52030.2-154          NPAGKEFMLESKYGGKLLGFHGRSG-EALDAIGPHFFAVNS
At1g52030.161-289        NVVGSKFVFE-ETSFKLVGFCGRSG-EAIDALGAHF-----
At1g52030.336-476        METEKKLELKDGKGGKLVGFHGKAS-DVLYALGAYFA----
At3g16400.2-142          KRPGVKFVL---HGGKIVGFHGRST-DVLHSLGAYVS----
At3g16440.2-144          LDTENKFVLKEKNGGKLVGFHGRAG-EILYALGAYF-----
At3g16440.154-300        IEAGTAFELKE-EGCKIVGFHGKVS-AVLHQFGVHILPVTN
At3g16470.2-145          LTSGEEAELG---GGKIVGFHGSSS-DLIHSVGVYIIPST-
At3g16470.158-297        LTSGEEAELG---GGKIVGFHGTSS-DLIHSLGAYIIP---
At3g16470.308-450        LEGGTEFVLEK-KDHKIVGFYGQAG-EYLYKLGVNVAPIA-
At3g21380.7-130          NKTRNQFSIHAPKDNQIAGFQGISS-NVLNSIDVHFA----
```

Fig. 6. Alignment of MyroBP-related sequences. Sequences of the N- and C-terminal domains of At3g16450.1 are aligned with those of MyroBP from *B. napus* and MyroBP-like proteins from *A. thaliana* (At1g52030, At3g16400, At3g16440, At3g16470 and At3g21380). Asterisks and vertical bars indicate identical and similar residues, respectively. The β-strands of At3g16450.1 are indicated by arrows above the sequence.

and ^{15}N-edited NOESY spectra. ^{15}N- and ^{13}C-edited NOESY spectra were recorded with a mixing time of 75 ms, and the inter-proton distance constraints were obtained from the NOESY peaks, which were selected and manually filtered using SPARKY.

Collection of conformational constraints, structure calculation and refinement

Automated NOE cross-peak assignments [30] and structure calculations with torsion-angle dynamics were performed using the program CYANA, version 2.2 [31]. Backbone torsion-angle constraints obtained from database searches using the program TALOS [16] were incorporated into the structure calculation. Simulated annealing with 20 000 torsion-angle dynamics time steps per conformer was performed during the CYANA structure calculations. In the final cycle of the CYANA protocol, 100 conformers were generated and further refined using the AMBER 9 software package [32] with a full-atom force field [33]. The refinement comprised three stages: initial minimization, molecular dynamics, and final minimization. Minimization and molecular dynamics consisted of 1500 steps and 20 ps duration, respectively. A generalized Born implicit solvent model was used to account for the solvent effects [34]. The force constants for distance and torsion-angle constraints were 50 kcal·mol^{-1}·Å^{-2} and 200 kcal·mol^{-1}·rad^{-2} respectively. From the resulting structures of this first AMBER refinement, we extracted backbone hydrogen-bond constraints in the regular secondary elements that were present in more than 75% of the 100 conformers. With these as additional constraints, we repeated the refinement. From the conformers that did not significantly violate experimental constraints, we selected the 20 lowest-energy structures for analysis. The structural quality was evaluated using PROCHECK-NMR [35]. The program MOLMOL [36] was used to visualize the structures. The coordinates of the 20 energy-refined CYANA conformers of At3g16450.1 have been deposited in the Protein Data Bank (accession code 2JZ4). The chemical shifts of At3g16450.1 have been deposited in the BioMagResBank (accession code 15607).

Frontal affinity chromatography

M9.1, 210.1, 210.4 and 210.1FX were purchased from Seikagaku Kogyo Co (Tokyo, Japan). The code numbers and structures of pyridylaminated oligosaccharides refer to the GALAXY website at http://www.glycoanalysis.info/ENG/index.html [37]. Two kinds of PA-oligosaccharides, GalNAcα1-3(Fucα1-2)Galβ1-3(Fucα1-4)GlcNAcβ1-3Galβ1-4Glc-PA and Neu5Acα2-6Galβ1-4GlcNAcβ1-2Manα1-6(Neu5Acα2-3Galβ1-3(Neu5Acα2-6)GlcNAcβ1-4(Neu5Acα2-6Galβ1-4GlcNAcβ1-2)Manα1-3)Manβ1-4GlcNAcβ1-4GlcNAc-PA were obtained from Takara Bio. Inc. (Otsu, Shiga, Japan). Other PA glycans were prepared by amination of the commercial oligosaccharides using 2-aminopyridine [38]. Lewis A- and Lewis X-type glycans, Galβ1-3(Fucα1-4)GlcNAcβ1-3Galβ1-4Glc and Galβ1-4(Fucα1-3)GlcNAcβ1-3Galβ1-4Glc were purchased from Calbiochem (San Diego, CA, USA). Cellohesaose, chitohesaose, isomaltohexaose, laminarihesaose and maltohexaose were purchased from Seikagaku Kogyo Co.

The protein At3g16450.1 containing the N-terminal histidine tag was dissolved in 10 mM HEPES buffer, pH 7.6, containing 150 mM NaCl, 1 mM $CaCl_2$, and bound to Ni-NTA agarose. After immobilization, the agarose beads were packed into a stainless steel column (4.0 × 10 mm, GL Sciences, Tokyo, Japan).

Frontal affinity chromatography analysis was performed as described previously [39]. PA oligosaccharides were dissolved at a concentration of 10 nM in 10 mM HEPES, pH 7.6, containing 150 mM NaCl, 1 mM $CaCl_2$, and applied onto the At3g16450.1 column at a flow rate of 0.25 mL·min^{-1} at 20 °C. The elution profile was monitored by the fluorescence intensity at 400 nm (excitation at 320 nm). Tetrasialyl PA glycan Neu5Acα2-6Galβ1-4GlcNAcβ1-2Manα1-6(Neu5-Acα2-3Galβ1-3(Neu5Acα2-6)GlcNAcβ1-4(Neu5Acα2-6Galβ1-4GlcNAcβ1-2)Manα1-3)Manβ1-4GlcNAcβ1-4GlcNA-PA was used as a control sugar to determine the elution volume of the unbound oligosaccharide.

NMR chemical-shift perturbation mapping

NMR samples were prepared using free [U-^{15}N]-labeled At3g16450.1 (0.1 mM protein, 10 mM HEPES, pH 7.6, 150 mM KCl, 1 mM $CaCl_2$) and its complex with PA sugar [same solvent composition plus 0.5 mM PA-(Glcα1-4Glc)$_3$]. ^{1}H-^{15}N HSQC spectra of the isolated and titrated samples were acquired at 27.5 °C using a Bruker Avance 600 MHz NMR spectrometer.

Acknowledgements

This work was supported by the Technology Development for Protein Analyses and Targeted Protein Research Program of the Ministry of Education, Culture, Sports, Science and Technology of Japan (MEXT), by Core Research for Evolutional Science and Technology (CREST) of the Japan Science and Technology Agency (JST), by a Grant-in-Aid for Scientific Research from the Japan Society for the Promotion of Science (JSPS), by the National Institutes of Health Protein Structure Initiative (grants P50 GM64598 and U54 GM074901), and by the Volkswagen Foundation.

References

1 The Arabidopsis Initiative (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* **408**, 796–815.

2 Vinarov DA, Loushin Newman CL & Markley JL (2006) Wheat germ cell-free platform for eukaryotic protein production. *FEBS J* **273**, 4160–4169.

3 Kainosho M, Torizawa T, Iwashita Y, Terauchi T, Ono AM & Güntert P (2006) Optimal isotope labelling for NMR protein structure determinations. *Nature* **440**, 52–57.

4 Rask L, Andréasson E, Ekbom B, Eriksson S, Pontoppidan B & Meijer J (2000) Myrosinase: gene family

evolution and herbivore defense in Brassicaceae. *Plant Mol Biol* **42**, 93–113.

5 Lönnerdal B & Janson JC (1973) Studies on myrosinases. II. Purification and characterization of a myrosinase from rapeseed (*Brassica napus* L.). *Biochim Biophys Acta* **315**, 421–429.

6 Xue J, Jørgensen M, Pihlgren U & Rask L (1995) The myrosinase gene family in *Arabidopsis thaliana*: gene organization, expression and evolution. *Plant Mol Biol* **27**, 911–922.

7 Takechi K, Sakamoto W, Utsugi S, Murata M & Motoyoshi F (1999) Characterization of a flower-specific gene encoding a putative myrosinase binding protein in *Arabidopsis thaliana*. *Plant Cell Physiol* **40**, 1287–1296.

8 Capella AN, Menossi M, Arruda P & Benedetti CE (2001) COI1 affects myrosinase activity and controls the expression of two flower-specific myrosinase-binding protein homologues in Arabidopsis. *Planta* **213**, 691–699.

9 Eriksson S, Andréasson E, Ekbom B, Granér G, Pontoppidan B, Taipalensuu J, Zhang J, Rask L & Meijer J (2002) Complex formation of myrosinase isoenzymes in oilseed rape seeds are dependent on the presence of myrosinase-binding proteins. *Plant Physiol* **129**, 1592–1599.

10 Taipalensuu J, Eriksson S & Rask L (1997) The myrosinase-binding protein from *Brassica napus* seeds possesses lectin activity and has a highly similar vegetatively expressed wound-inducible counterpart. *Eur J Biochem* **250**, 680–688.

11 Falk A, Taipalensuu J, Ek B, Lenman M & Rask L (1995) Characterization of rapeseed myrosinase-binding protein. *Planta* **195**, 387–395.

12 Kasai K, Oda Y, Nishikawa M & Ishii S (1986) Frontal affinity chromatography: theory for its application to studies on specific interactions of biomolecules. *J Chromatogr* **376**, 33–47.

13 Sugimori N, Torizawa T, Aceti DJ, Thao S, Markley JL & Kainosho M (2004) ^{1}H, ^{13}C and ^{15}N backbone assignment of a 32 kDa hypothetical protein from *Arabidopsis thaliana*, At3g16450.1. *J Biomol NMR* **30**, 357–358.

14 Cavanagh J, Fairbrother WJ, Palmer AG III, Skelton NJ & Rance M (2006) *Protein NMR Spectroscopy. Principles and Practice*, 2nd edn. Academic Press, San Diego, CA.

15 Seavey BR, Farr EA, Westler WM & Markley JL (1991) A relational database for sequence-specific protein NMR data. *J Biomol NMR* **1**, 217–236.

16 Cornilescu G, Delaglio F & Bax A (1999) Protein backbone angle restraints from searching a database for chemical shift and sequence homology. *J Biomol NMR* **13**, 289–302.

17 Güntert P, Mumenthaler C & Wüthrich K (1997) Torsion angle dynamics for NMR structure calculation with the new program DYANA. *J Mol Biol* **273**, 283–298.

18 Güntert P (2003) Automated NMR protein structure calculation. *Prog Nucl Magn Reson Spectrosc* **43**, 105–125.

19 Kabsch W & Sander C (1983) Dictionary of protein secondary structure – pattern-recognition of hydrogen-bonded and geometrical features. *Biopolymers* **22**, 2577–2637.

20 Holm L, Ouzounis C, Sander C, Tuparev G & Vriend G (1992) A database of protein structure families with common folding motifs. *Protein Sci* **1**, 1691–1698.

21 Holm L & Sander C (1993) Protein structure comparison by alignment of distance matrices. *J Mol Biol* **233**, 123–138.

22 Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W & Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25**, 3389–3402.

23 Falk A, Ek B & Rask L (1995) Characterization of a new myrosinase in *Brassica napus*. *Plant Mol Biol* **27**, 863–874.

24 Schuler GD (1997) Pieces of the puzzle: expressed sequence tags and the catalog of human genes. *J Mol Med* **75**, 694–698.

25 Takeda M, Ikeya T, Güntert P & Kainosho M (2007) Automated structure determination of proteins with the SAIL-FLYA NMR method. *Nat Protoc* **2**, 2896–2902.

26 López-Méndez B & Güntert P (2006) Automated protein structure determination from NMR spectra. *J Am Chem Soc* **128**, 13112–13122.

27 Scott A, López-Méndez B & Güntert P (2006) Fully automated structure determinations of the Fes SH2 domain using different sets of NMR spectra. *Magn Reson Chem* **44**, S83–S88.

28 Torizawa T, Shimizu M, Taoka M, Miyano H & Kainosho M (2004) Efficient production of isotopically labeled proteins by cell-free synthesis: a practical protocol. *J Biomol NMR* **30**, 311–325.

29 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.

30 Herrmann T, Güntert P & Wüthrich K (2002) Protein NMR structure determination with automated NOE assignment using the new software CANDID and the torsion angle dynamics algorithm DYANA. *J Mol Biol* **319**, 209–227.

31 Güntert P (2004) Automated NMR structure calculation with CYANA. *Methods Mol Biol* **278**, 353–378.

32 Case DA, Cheatham TE, Darden T, Gohlke H, Luo R, Merz KM, Onufriev A, Simmerling C, Wang B & Woods RJ (2005) The Amber biomolecular simulation programs. *J Comput Chem* **26**, 1668–1688.

33 Cornell WD, Cieplak P, Bayly CI, Gould IR, Merz KM, Ferguson DM, Spellmeyer DC, Fox T, Caldwell JW & Kollman PA (1995) A second generation force

field for the simulation of proteins, nucleic acids, and organic molecules. *J Am Chem Soc* **117**, 5179–5197.

34 Tsui V & Case DA (2000) Theory and applications of the generalized Born solvation model in macromolecular simulations. *Biopolymers* **56**, 275–291.

35 Laskowski RA, Rullmann JAC, MacArthur MW, Kaptein R & Thornton JM (1996) AQUA and PROCHECK-NMR: programs for checking the quality of protein structures solved by NMR. *J Biomol NMR* **8**, 477–486.

36 Koradi R, Billeter M & Wüthrich K (1996) MOLMOL: a program for display and analysis of macromolecular structures. *J Mol Graphics* **14**, 51–55.

37 Takahashi N & Kato K (2003) GALAXY (glycoanalysis by the three axes of MS and chromatography): a web application that assists structural analyses of *N*-glycans. *Trends Glycosci Glycotechnol* **15**, 235–251.

38 Yamamoto S, Hase S, Fukuda S, Sano O & Ikenaka T (1989) Structures of the sugar chains of interferon-γ produced by human myelomonocyte cell line HBL-38. *J Biochem (Tokyo)* **105**, 547–555.

39 Arata Y, Hirabayashi J & Kasai K (2001) Sugar binding properties of the two lectin domains of the tandem repeat-type galectin LEC-1 (N32) of *Caenorhabditis elegans*. Detailed analysis by an improved frontal affinity chromatography method. *J Biol Chem* **276**, 3068–3077.