Journal of Biomolecular NMR (2006) 36:267–277
DOI 10.1007/s10858-006-9092-z

Article

Structure validation of the Josephin domain of ataxin-3: Conclusive evidence for an open conformation

Giuseppe Nicastro[a], Michael Habeck[b], Laura Masino[a], Dmitri I. Svergun[c,d] & Annalisa Pastore[a,*]

[a]*National Institute for Medical Research, The Ridgeway, London NW7 1AA, UK;* [b]*Max-Planck Institutes for Developmental Biology and for Biological Cybernetics, Spemannstr. 35–38, Tübingen 72076, Germany;* [c]*European Molecular Biology Laboratory Hamburg Outstation, Notkestrasse 85, Hamburg 22603, Germany;* [d]*Institute of Crystallography, Russian Academy of Sciences, Leninsky pr. 59, Moscow 117333, Russia*

Received 18 April 2006; Accepted 19 September 2006

Key words: cysteine protease, polyglutamine disease, SCA3, structure, ubiquitin, Ubl

Abstract

The availability of new and fast tools in structure determination has led to a more than exponential growth of the number of structures solved per year. It is therefore increasingly essential to assess the accuracy of the new structures by reliable approaches able to assist validation. Here, we discuss a specific example in which the use of different complementary techniques, which include Bayesian methods and small angle scattering, resulted essential for validating the two currently available structures of the Josephin domain of ataxin-3, a protein involved in the ubiquitin/proteasome pathway and responsible for neurodegenerative spinocerebellar ataxia of type 3. Taken together, our results demonstrate that only one of the two structures is compatible with the experimental information. Based on the high precision of our refined structure, we show that Josephin contains an open cleft which could be directly implicated in the interaction with polyubiquitin chains and other partners.

Introduction

The development of new tools for fast and efficient structure determination has permitted the rapid and apparently inexhaustible growth of the number of three-dimensional (3D) structures available in the protein database. Full advantage of the richness of these archives crucially depends on the quality and the reliability of the structures deposited. This requires the development of new approaches which may aid in the identification of possible human mistakes, give an estimate of the structure accuracy and provide altogether information on the degree of reliability of a given structure. Here, we show that it is possible, using the most advanced computational and experimental tools, to validate structures even when their differences are apparently subtle and difficult to assess.

We have used as a paradigmatic example our recently solved structure of the Josephin domain. Josephin is an N-terminal domain and the only constitutively folded region of ataxin-3, a human protein involved in the rare but dominant Joseph-Machado disease, also known as spinocerebellar ataxia of type 3 (SCA3) (Kawagushi et al., 1994; Taylor et al., 2002; Masino et al., 2003; Masino et al., 2004). Structure determination of Josephin by nuclear magnetic resonance (NMR) (PDB

*To whom correspondence should be addressed.
E-mail: apastor@nimr.mrc.ac.uk

entry 1yzb, Nicastro et al., 2005) proved that the domain has a predominantly α-helical fold typical of a cysteine protease, thus demonstrating that this function, known for the full-length protein, is localised in the N-terminus of the protein (Burnett and Pittman, 2003) (Figure 1). A feature that is specific of the Josephin fold is the presence of a helical hairpin which contains helices α2 and α3. Shortly after the first publication, another structure became available, also solved by NMR (2aga, Mao et al., 2005). Although in general agreement, the two structures differ significantly in at least two regions: the helical hairpin and the C-terminus. In one structure (1yzb), the hairpin region is flexible, as demonstrated by both the root mean square deviation (r.m.s.d) of the structure bundle and by the relaxation parameters, and protrudes out into solution, thus creating a cleft in which other ligands could insert. In the other structure (2aga), the hairpin seems stiff and packs against the rest of the globular domain. Assessing whether these differences are genuine or arise from the low resolution of NMR methods is of great functional importance, since the cleft is the region where ubiquitin and possibly other substrates are thought to interact.

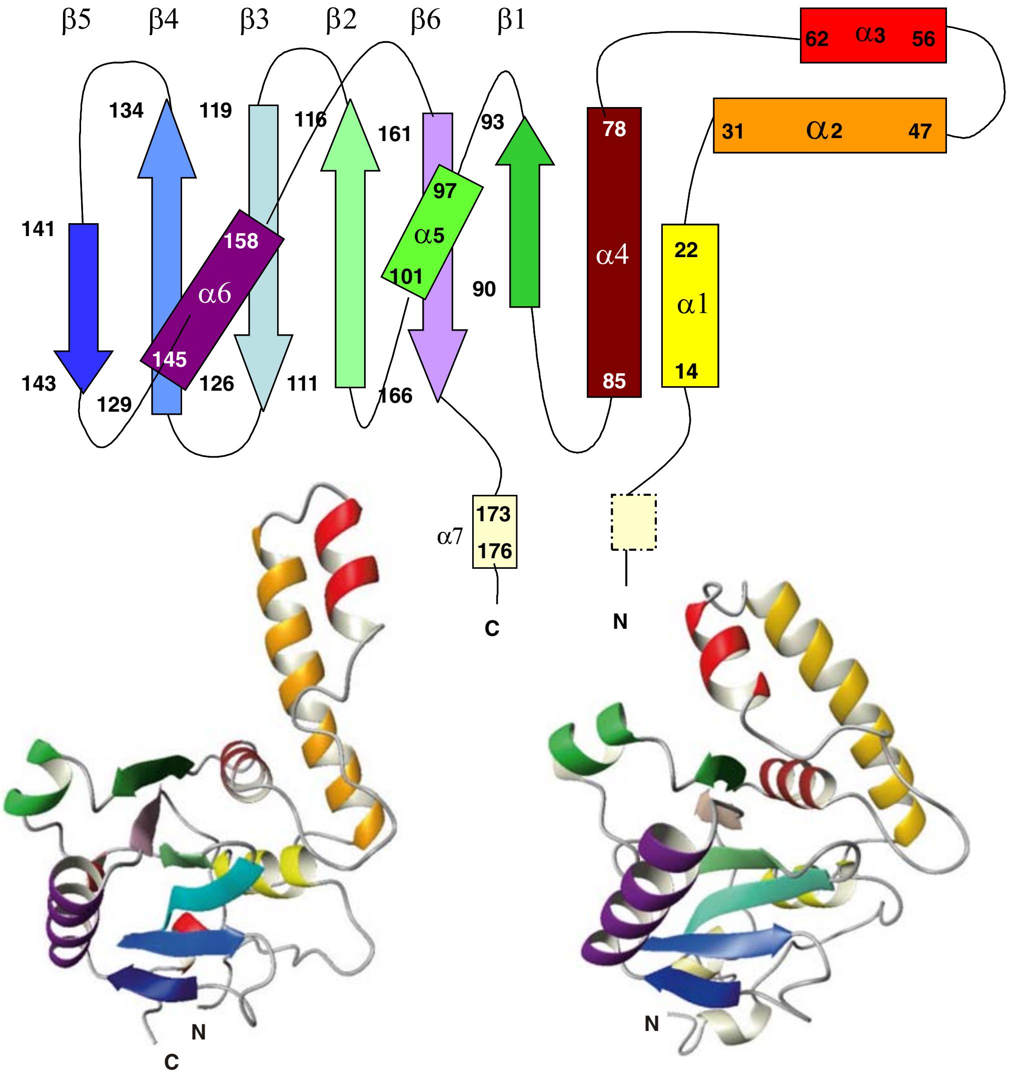

Figure 1. Comparison of the ribbon representations of 1yzb (left) and 2aga (right). The backbone atoms of the core residues were used the first fit and then one of the two structures was translated.

To compound the debate and as a prerequisite for further structural studies of ataxin-3, we have developed a combined strategy to validate the two structures. We have first used Bayesian methods (Habeck et al., 2005a; Rieping et al., 2005a) to estimate the internal consistency of the two structure bundles with the respective sets of distance restraints. This probabilistic approach, well known in other fields, has only recently been applied to NMR structure determination and provides a powerful way to assess the consistency of a structure with the original data in rigorous statistical terms. Here, we have pushed the limits of its application to a molecule of the size of the Josephin domain (182 residues). Residual dipolar couplings and quality control methods were then used to compare the two structures.

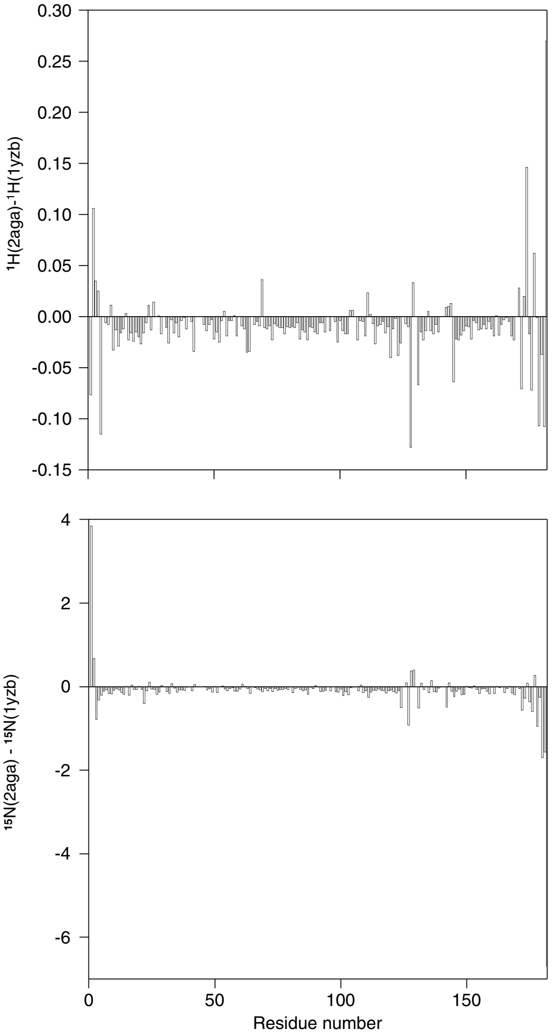

Figure 2. Differences of ^{1}H and ^{15}N chemical shifts reported for 1yzb and 2aga, respectively.

Finally, independent and conclusive validation came from small angle scattering (SAXS). This technique, although being a low resolution method, is highly sensitive not only to the overall shape but also to the internal structure of macromolecules and has been successfully used to validate high resolution crystallographic structures and predicted homology models in solution (see Svergun and Koch, 2003 for a review).

Taken together, our data provide a clear estimate of the accuracy of the Josephin structures and prove conclusively the presence of an open cleft between the hairpin and the main body of the domain. Because of its closeness to the active site, the cleft is likely to be directly involved in the cysteine protease activity of the domain and be responsible for accommodating poly-ubiquitin chains and other molecular partners. Our data thus provide a highly reliable description of the Josephin structure which may be used as a reference for further structural and functional studies. We believe that a similar strategy may prove valuable also in other structure validation.

Materials and methods

Protein production

The protein was produced as described elsewhere (Nicastro et al., 2004; Nicastro et al., 2005) and either used immediately after purification or frozen to −20 °C. All the samples were filtered, prior to use using a 0.22 μm membrane to eliminate possible large aggregates.

NMR measurements

Spectra were recorded on a Varian spectrometers operating either at 600 or 800 MHz at 25 °C. Josephin samples in 20 mM sodium phosphate buffer at pH 6.5 were used. RDCs were measured in polyacrilamide gels (Sass et al., 2000). Alternative methods were attempted but proved unsuccessful due to the intrinsic tendency of Josephin to aggregate, further enhanced by the confinement effect in liquid crystalline media (Munishkina et al., 2004). 38 RDCs were nonetheless obtained.

Bayesian structure calculations

A more comprehensive explanation of the method can be found in the relevant literature. In short, a Bayesian structure calculation determines both the coordinates X and the error σ (or weight in the traditional view) of the dataset. The unknown structure is represented through a conditional probability $p(X) = \mathrm{d}P(X|D, I)/\mathrm{d}X$ that quantifies the likelihood that X is the 'true' molecular structure given the dataset D and other relevant prior knowledge I (Habeck et al. 2005a; Rieping et al., 2005a). The posterior distribution $p(X)$ spreads the uncertainty about the structure over the entire conformational space and peaks in regions where conformations are in agreement with the data and the prior knowledge. The posterior distribution is proportional to the product of the likelihood function $L(X)$ and the prior distribution $\pi(X)$: $p(X) \infty L(X)\ \pi(X)$. The likelihood function derives from the probability of observing the measurements given the molecular structure, i.e., $L(X) = P(D|X, I)$. In case of NOE data the lognormal distribution with unknown width σ is an appropriate likelihood function (Rieping et al., 2005b).

The errors serve as an objective figure of merit to evaluate the quality of the data (note that for a lognormal model the error has no units). Roughly speaking, large errors (i.e. σ greater than 1.5) indicate low quality data which possibly contain distances from erroneously assigned cross peaks or from peaks that are strongly affected by protein dynamics or spin diffusion. Small errors (i.e. σ smaller than 0.9) indicate data that are highly consistent internally and with the prior knowledge encoded in the force field. Typical values for σ using the lognormal model range from 0.9 to 1.3 (Rieping et al., 2005b). A more advanced likelihood function (W. Rieping, M. Habeck, in preparation) allows us to automatically downweight peaks that are found to be inconsistent with the rest of the data set, thus lowering the estimated error of the remaining subset of consistent restraints. Internal inconsistencies typically originate from assignment errors.

The 1yzb dataset comprises 5525 unambiguous and 925 ambiguous distances. The data set was augmented with 44 additional hydrogen bond restraints to allow for a fair comparison with the published structure 1yzb and to improve the convergence of the structure calculations. Residual dipolar couplings were not included in the calculations but used for a posteriori validation.

Two structure calculations were run starting from an extended conformation and using 50 replicas, which were simulated in parallel at different generalized temperatures (Habeck et al., 2005b). The first run uses likelihood model I, which is based on the lognormal distribution, to analyze the distance data, the second is based on the extended likelihood model II, which automatically downweights inconsistent data. We combined various Markov chain Monte Carlo methods to investigate the high-dimensional joint probability distribution of the coordinates and of the errors. Notably, we used a multi-parameter replica-exchange Monte Carlo scheme (Habeck et al., 2005b), which circumvents trapping of the Markov chain in single modes and guarantees full sampling of the most likely conformations and errors.

SAXS measurements

The synchrotron radiation X-ray scattering data were collected on the X33 camera (Koch and Bordas, 1983; Boulin et al., 1988) of the EMBL (DORIS III, DESY). Solutions of Josephin were measured at 12 °C and at protein concentrations of 3.8, 9.8 and 16 mg/ml using a setup with two proportional gas detectors (Gabriel and Dauvergne, 1982). At the two sample-detector distances of 1.1 and 2.7 m and a wavelength $\lambda = 1.5$ Å the total range of momentum transfer covered was $0.015 < s < 0.95$ Å^{-1} ($s = 4\pi \sin(\theta)/\lambda$ where θ is the scattering angle). To check for radiation damage, the data were collected in 15 successive 1-minute frames. The data were averaged after normalization to the intensity of the incident beam, corrected for the detector response and the scattering of the buffer was subtracted. The difference data were extrapolated to zero solute concentration following standard procedures and the curves measured in different angular intervals were merged. Protein concentration was determined by UV absorption, assuming a calculated extension coefficient at 280 nm of 24750 M^{-1}cm^{-1}. Note that for the extrapolation to zero concentration, it is only required that the relative and not the absolute concentrations are correct. This implies that a systematic error in the assumption of the OD plays no role and only errors in the actual absorption measurements matter. These errors do not exceed 5%. All data manipulations were performed using the program package PRIMUS (Konarev et al., 2003).

The forward scattering $I(0)$, radius of gyration R_g and the maximum particle dimension $D_{\max}$. were evaluated using the indirect transformation program GNOM (Svergun, 1992). The molecular mass of the solute was evaluated by comparison of the forward scattering with that from reference solutions of bovine serum albumin (which has a molecular mass of 66 kDa). The excluded volume of the hydrated particle was computed from the small angle portion of the data ($s < 0.33$ Å^{-1}) using the equation (Porod, 1982):

$$V = 2\pi^2 I(0) / \int_0^\infty s^2 I_{\exp}(s) ds \tag{1}$$

Prior to this analysis an appropriate constant was subtracted from each data point to force the s^{-4} decay of the intensity at higher angles following the Porod's law (Porod, 1982) for homogeneous particles. This ''shape scattering'' curve was further used to generate the low resolution *ab initio* model of Josephin the program DAMMIN, (Svergun, 1999) which represents the protein by an assembly of compact interconnected beads. An alternative higher resolution *ab initio* model was constructed using the full range of scattering data by the program GASBOR (Svergun et al., 2001) representing the protein as an assembly of dummy residues. The scattering from the NMR models of Josephin was calculated using the program CRYSOL (Svergun et al., 1995), which adjusts the excluded volume of the particle and the contrast of its hydration layer to fit the experimental data $I_{\exp}(s)$ to minimize discrepancy:

$$\chi^2 = \frac{1}{N-1}\sum_j \left[\frac{I_{\exp}(s_j) - cI_{calc}(s_j)}{\sigma(s_j)}\right]^2 \tag{2}$$

where N is the number of experimental points, c is a scaling factor and $I_{\text{calc}}(s)$ and $\sigma(s_j)$ are the calculated intensity and the experimental error at the momentum transfer s_j, respectively.

The results of multiple *ab initio* runs were analyzed and compared with the NMR structures of Josephin using the programs DAMAVER (Volkov and Svergun, 2003) and SUPCOMB (Kozin and Svergun, 2001). The latter program aligns two arbitrary low or high resolution models represented by ensembles of points by minimizing a dissimilarity measure called normalized spatial discrepancy (NSD). For every point (bead or atom) in the first model, the minimum value among the distances between this point and all points in the second model is found, and the same is done for the points in the second model. These distances are added and normalized against the average distances between the neighboring points for the two models. Generally, NSD values close to unity indicate that the two models are similar. The program DAMAVER generates the average model of the set of superimposed structures and also specifies the most typical model (i.e. that having the lowest average NSD with all the other models of the set).

Results

Degree of similarity of the two constructs

The difference between the two structures could in principle arise from a genuine difference in the samples, due to specific purification protocols or experimental conditions. To assess the importance of these factors, we compared the reported chemical shifts (bmrb entries 6241 and 6742, Nicastro et al., 2004; Mao et al., 2005) used as the basis for structure calculations (Figure 2). Chemical shifts are parameters extremely sensitive to the chemical environment and would therefore reflect any difference between the two constructs. The backbone amide chemical shifts are in excellent agreement, with a $<\Delta\delta>$ of less than 0.02 ppm and 0.1 ppm in the proton and nitrogen dimensions respectively, with the only exception of the N- and C-terminal residues. The agreement is specifically high in the region which spans the $\alpha2/\alpha3$ helical hairpin (residues 30–65), whereas the main differences are between residues 120–145 which correspond to regions far away from the hairpin. A similar agreement is also observed for the chemical shifts of the side chains (data not shown). This comparison indicates that it is highly unlikely that the resulting structures could have significant intrinsic differences.

Structure validation by assessing consistency with the relative datasets by Bayesian methods

Before any further comparison between the structures, we validated 1yzb by recalculating it using Bayesian methods (Rieping et al., 2005a, b; Habeck et al., 2005a, b). This approach, which is based on a probabilistic framework, represents the uncertainty about a structure by a probability distribution over the conformational space. The shape of the distribution quantifies objectively the precision and uniqueness of the structure. The data reliability is estimated during the structure calculation and quantified by an error, which determines the weight of a single dataset relative to a force field and to other data sets. The major advantages of a Bayesian approach are that the obtained structure ensembles are statistically meaningful and that it is possible to estimate the mutual consistency of additional unknown parameters such as weighting constants, calibration factors, and Karplus' parameters directly from the data. However, the richness of the information obtained comes at the price of elevated computational costs, which make its application increasingly challenging with the molecule size. Josephin constitutes the first application of the method to a system of this size (182 residues).

The 1yzb was evaluated by two different likelihood functions, which represent the probability of observing a specific molecular structure given a set of measurements. Of these models, the first one (model I) describes NOE-based distances assuming the isolated spin pair approximation (ISPA) to predict the intensities from the structure and relies on a lognormal distribution (Rieping et al., 2005b) with an unknown error that is estimated using posterior sampling (Habeck et al., 2006). The second likelihood function (model II) constitutes an extended version of model I (W. Rieping and M. Habeck, in preparation) and allows one to identify peaks that are inconsistent with the rest of the dataset.

When analysing the 1yzb dataset with likelihood model I, we obtained an ensemble of structures with an exposed hairpin, which confirms the previously published structure (Figure 3). The precision of the structure core is very high, with a local r.m.s.d. ranging from 0.2 Å to 0.3 Å. The Bayesian ensemble is even tighter than in the original bundle. Most of the conformational variability is observed in the helical hairpin with a local r.m.s.d. rising to 1.4 Å. The estimated error of the data is $\sigma = 1.14 \pm 0.01$ and $\sigma = 1.35 \pm 0.03$ for the unambiguous and ambiguous distances, respectively. These values are well within those typical for NOESY data (Rieping et al., 2005b). The larger error in the ambiguous distances is likely caused by small inconsistencies in the ambiguous assignments, which are more difficult to control than unambiguous data. This indicates that the derived structure ensemble has a relatively high level of reliability.

To investigate further the internal consistency of the datasets and to identify erroneous data that could be responsible for the extended conformation of the hairpin we applied likelihood model II to the same data. In this way, we can estimate the number and the distribution of internally consistent restraints. These are 91 ± 1% of the unambiguous distances and 88 ± 2% of the ambiguous restraints, according to the observation that ambiguous assignments are more likely to contain small inconsistencies. These however involve mostly intra-residue and sequential restraints which are not crucial for the topology of the structure and are not preferably located in or near the hairpin region but scattered all over the structure. The percentage of internally consistent distances is anyway very high in both sets of restraints. Most of the inconsistencies had been filtered out during the previous ARIA calculations so that the resulting 1yzb structure bundle would not reflect their presence. If only consistent restraints are considered, the relative error decreases to $\sigma = 0.88 \pm 0.01$ and $\sigma = 0.97 \pm 0.04$ for the unambiguous and ambiguous data, respectively. Comparison of the conformational structure bundles obtained when adopting each of the two models shows that they are highly similar, with the only exception for the hairpin variability which is slightly larger in model II with a local r.m.s.d. of up to 2 Å (Figure 3). This indicates that the inconsistent restraints identified with model II are not responsible for the overall fold and confirm the validity of the data used for structure determination of 1yzb.

Structure validation of 2aga and 1yzb

Once established the reliability of the 1yzb bundle, we compared and cross-validated the 1yzb and 2aga structures according to different independent criteria. First, we evaluated the quality of the two structure bundles using standard quality control methods. Several packages have been developed in the last two decades to assist structure validation.

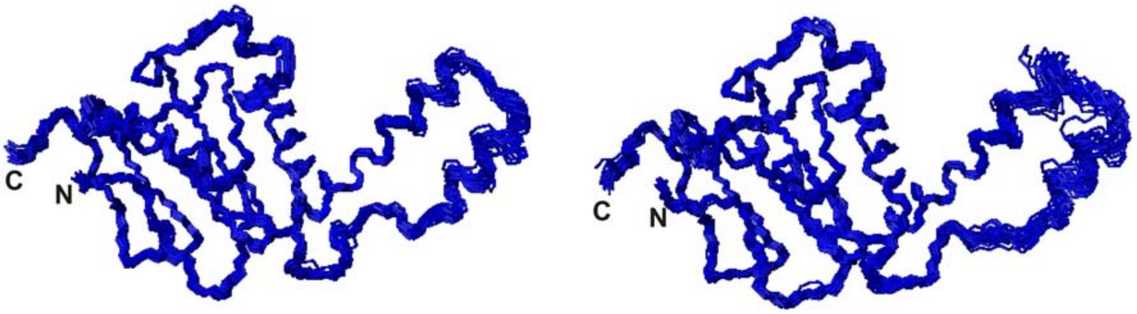

Figure 3. Superposition of the backbone atoms structure ensembles generated by Bayesian methods as generated from likelihood model I using replica Monte Carlo (left) and model II which is based on an extended likelihood model (right).

The most commonly used is probably the statistics of the Ramachandran plot (Laskowski et al., 1996). However, while an acceptable Ramachandran map is a "*conditio sine qua non*" for considering a structure acceptable, this test is highly insensitive to poor side chain packing or to distorted geometries. We therefore compared the Procheck analysis (Lakowski et al., 1996) of the two NMR bundles with the packing parameters as calculated by the WhatIf program (Vriend, 1990), one of the major quality control tools. The final statistics are shown in Table 1. It is clear that, while no major differences are observed in the Procheck analysis, the 2aga bundle has a geometry and side chain packing quality noticeably inferior to that of 1yzb.

As another powerful tool to validate the structure, we measured RDC values and compared the experimental values with those predicted on the basis of each of the two structure bundles. Because of its ability to determine the orientation of a molecule or of parts of it respect to an external reference system, this approach has been successfully used in the past for solving several structural discrepancies (Bax, 2003). We had on purpose excluded RDCs from the Bayesian calculations to be able to use them as independent parameters. Their measurement was however not easy for a protein such as Josephin which has a strong tendency to aggregate, since this property is enhanced in the confining media usually used for alignment, such as gels or bicelles.

We eventually managed to obtain 38 RDCs, which had previously been included in the structure calculation of 1yzb (Nicastro et al., 2005). Comparison of the experimental values with those calculated for each of the two structures shows that 1yzb is in excellent agreement with the data, having an average difference of 0.57 (Figure 4). Conversely, practically no correlation is observed for the 2aga structure, thus supporting the accuracy of the 1yzb structure.

SAXS studies support an elongated shape of Josephin

Finally, we used SAXS measurements to have an independent description of the overall shape of Josephin. They revealed that concentrated

Table 1. Quality control indices of the Josephin structures as calculated with two of the major programs for structure validation. For a structure to be acceptable, the WhatIf quality indices should be as more positive as possible. Values consistently below −3 are usually symptomatic of a wrong structure

	1yzb	2aga
Whatif quality check		
First generation packing quality	−1.108	−1.923
Second generation packing quality	−1.726	−3.453
Ramachandran plot appearance[a]	−3.837	−2.705
χ^1–χ^2 rotamer normality	−2.267	−5.284
Backbone conformation	−0.799	−4.649
Procheck Ramachandran statistics (%)		
Most favoured region	85.5	78
Additional allowed regions	12.8	12.8
Generously allowed regions	0.7	5.8
Disallowed regions	1	3.4

[a] This number is the sum of the score obtained for each amino acid which can range from 0 to 1.0. Due to the tighter CNS force field (Brunger et al., 1998), used to calculate 1yzb, this structure has several small violations which lower the score but very few residues in the disallowed region. The 2aga structure viceversa, which was calculated with the CYANA forcefield (Mao et al., 2005), has more residues in the disallowed region but the ones in the allowed region score higher.

solutions of Josephin have a strong tendency to form unspecific aggregates, leading to an artificial increase of the scattered intensity at very low angles. This is consistent with what was observed in previous studies, which had shown that the Josephin domain has a strong tendency to aggregate and form fibres (Masino et al., 2004; Nicastro et al., 2005). NMR relaxation studies had however established that, when used immediately after purification or fast thawing, the largely predominant species in solution is the monomer (Nicastro et al., 2005). At the temperature and concentrations used for the SAXS measurements, we estimate that the kinetics of aggregation would have a half-life time of more than 48 h, that is well above the time needed for SAXS measurements (Masino et al., 2004). The effect of a minor population of aggregate in the sample used in the SAXS measurements could be reliably removed by extrapolation to zero concentration as illustrated in the insert of Figure 5A. The extrapolated data display a linear Guinier plot further suggesting that the aggregation effects are removed (Figure 5B). This procedure resulted in a molecular mass of the solute, as estimated by extrapolating the intensity to zero angle (23 ± 2 kDa), compatible with that of monomeric Josephin, thus proving that the aggregation effects were removed. This finding is further corroborated by the excluded volume of the particle in solution, which is equal to (35 ± 5) × 10^3 Å^3. It should be borne in mind that, for small globular proteins, the hydrated volume in Å^3 should be about 1.5–2 times the molecular mass in Da units. The experimental radius of gyration R_g and maximum size D_{max} (20.0 ± 0.5 Å and 65 ± 10 Å, respectively) suggested an elongated particle shape.

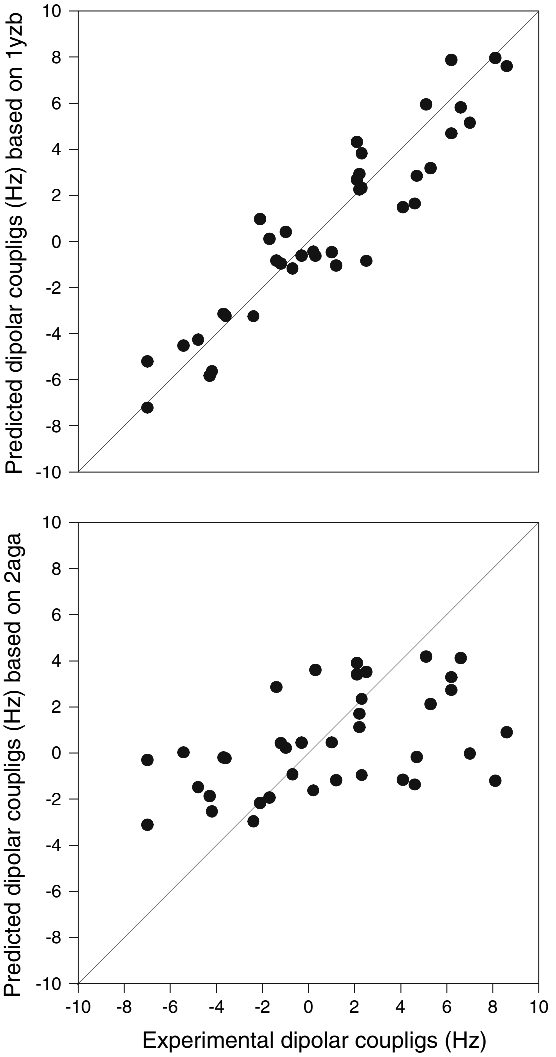

Figure 4. Comparison of experimental ^{1}HN–^{15}N dipolar couplings with the values predicted from the 1yzb (average difference between experimental and calculated values is 0.57) (A) and 2aga (average difference is 3.37) (B) structures.

The low resolution shape of Josephin was reconstructed *ab initio* using the bead modelling program DAMMIN by fitting the scattering data up to about 20 Å resolution. More detailed *ab initio* models were built by the dummy residues program GASBOR, using the full data range. Several independent simulated annealing reconstructions using the two programs yielded reproducible results neatly fitting the experimental scattering data (discrepancy χ equal to 1.06 and 0.86 for DAMMIN and GASBOR, respectively; fits not shown). Superposition of the most probable low resolution model of Josephin constructed by GASBOR onto the two NMR models shows that 1yzb agrees better with the SAXS models than 2aga (Figure 5C). The *ab initio* models yield also quantitatively a better agreement with the NMR structure 1yzb (NSD = 1.04 and 1.06 for DAMMIN and GASBOR models, respectively), than with 2aga (NSD = 1.11 and 1.26 for DAMMIN and GASBOR models, respectively).

As a further validation of the two NMR models, their scattering patterns were computed and compared with the experimental SAXS data (Figure 5A). 1yzb displays once again a better fit to the experiment (χ = 1.18) than 2aga (χ = 1.35). It must be noted that the larger discrepancy of 2aga arises not only from the low angle scattering region, but also from the range around s = 0.2 Å^{-1}, i.e. resolution about 30 Å, which reflects solely the overall internal structure. Moreover, the distance distribution function $P(r)$ computed from the experimental data agrees much better with the distribution of 1yzb than with that of 2aga (Figure 5D). Therefore, all comparisons indicate that 1yzb is significantly more consistent with the SAXS data than 2aga.

Discussion

The Josephin domain is an essential functional region of ataxin−3, a protein component of the ubiquitin proteasome pathway able to bind and cleave polyubiquitin chains containing four or more ubiquitins (Kawagushi et al., 1994; Donaldson et al., 2003; Burnett and Pittman, 2003; Chai et al., 2004). The cysteine protease activity observed for the full-length protein has previously been mapped within the Josephin domain, which was shown to be able to cleave ubiquitin substrates and bind the specific ubiquitin protease inhibitor, ubiquitin-aldehyde (Burnett and Pittman, 2003; Nicastro et al., 2005). It is therefore important to have a reliable structure of Josephin which could be used both in mutant design and as a basis for further structural studies. A correct definition of the overall shape of the molecule and of its surface is also important for mapping its interactions with other cellular components.

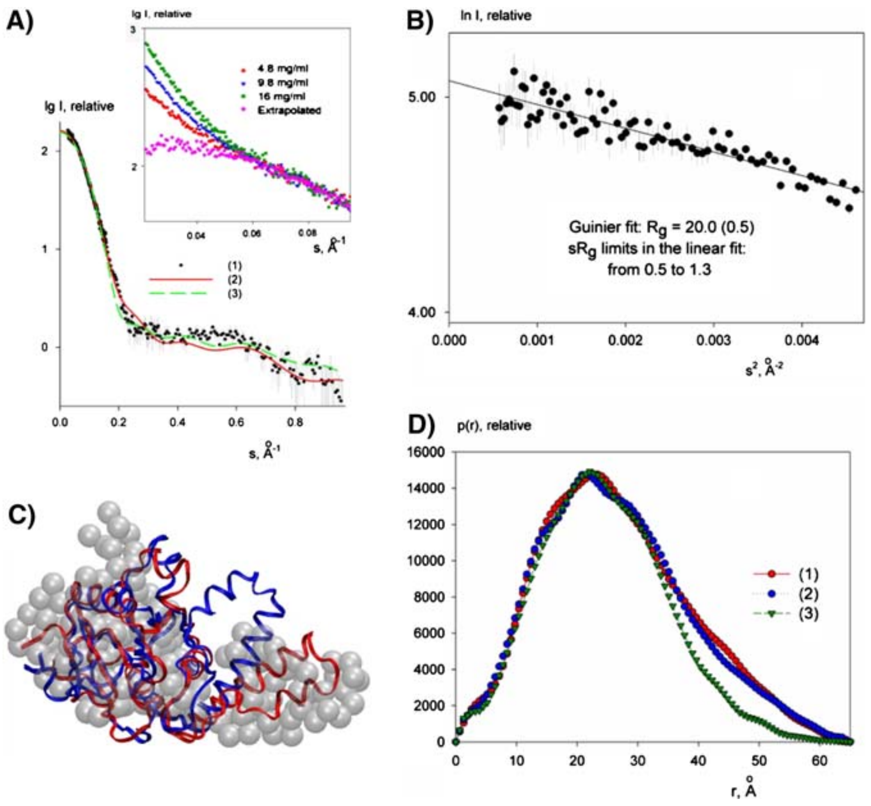

Figure 5. SAXS analysis and comparison with the two NMR structure. (A) Plot of the logarithm of the scattering intensity versus the momentum transfer (in Å^{-1}). The experimental SAXS data with error bars (1) are compared with the scattering computed from the NMR structures (1yzb (2) and 2aga (3), respectively). Extrapolation to zero solute concentration to remove the aggregation effects is illustrated in the insert. (B) Guinier plot of the data extrapolated to zero concentration (the straight line corresponds to the fit); (C) Overlay of the NMR models displayed in a ribbon representation (red for 1yzb, and blue for 2aga) with the most probable ab initio model of Josephin as generated by the program GASBOR (semi-transparent beads represent the Cα positions of dummy residues). (D) distance distribution functions (the numbering as in Figure 5B).

Here, we have used a combination of methods to evaluate the quality, accuracy and mutual consistency of the two structures currently available for the Josephin domain (Mao et al., 2005; Nicastro et al., 2005). It is interesting to learn from this example, the possible limitations which restrict the use of the techniques applied here and, at the same time, consider whether our approach bears sufficient generality to be applied more widely. We first used Bayesian methods to estimate the degree of reliability of 1yzb and its consistency with the experimental restraints which it is based on. In doing so, we have pushed the molecular size affordable for application of Bayesian methods to proteins. An interesting result of our work is that calculations starting with the published structure produced similar results for the estimated errors to those calculated after a full *de novo* structure determination. The 1yzb dataset turned out to be overall highly self-consistent, with errors well within those obtained for other test examples (Rieping et al., 2005b; Habeck et al. 2006). Bayesian methods, only relatively recently exploited in NMR structure calculations, will most likely become increasingly used as a new tool for structure validation. However, while very useful in discriminating data inconsistencies, they might remain inconclusive when having to distinguish between datasets of restraints and structures each individually consistent internally.

When we then assessed the 1yzb and 2aga datasets in terms of their agreement with a set of RDCs, only 1yzb showed an excellent agreement with the experimental data, suggesting that this structure has higher accuracy. RDCs are undoubtedly a powerful tool for structure refinement which has been widely used for structure validation. Because working with a protein with high tendency to aggregate, we have however encountered one of the few limitations of this approach: the molecular crowding effect induced by most of the agents used for sample alignment is due to fasten aggregation and, depending on the kinetics, prevent or seriously limits recording of a sufficient set of RDC restraints. The lower quality of 2aga was also evaluated by one of the standard, although still seldom used in structure papers, programs for structure validation (Vriend, 1990). Our results strongly suggest that analysis with this or equivalent packages should always be presented in structural papers, as they constitute, together with the Ramachandran plot, important and complementary criteria for structure evaluation.

Finally, we used SAXS studies to determine independently the Josephin shape. Although the use of SAXS techniques may be useful only if there are significant changes in the overall structure, they can provide important information and are now used routinely to validate crystal structures in solution (see for instance Vestergaard et al., 2005). We had in facts resorted to SAXS well before solving the structure of Josephin as a way to study aggregation and have a preliminary low resolution picture of the domain. The data were collected and the low resolution model in Figure 5 was obtained in 2003 but left unpublished until the problem of structure validation arose. It is now clear that the SAXS data are fully in agreement with the other evidence and add an independent and complementary validation of the Josephin structure.

Overall, we must conclude that, under the published experimental conditions, Josephin is present in solution in an open semi-elongated L-shape conformation. The helical hairpin, not being tightly packed against the rest of the structure is relatively more mobile and able to produce low-frequency motions around an equilibrium point, as reflected by the relaxation parameters published elsewhere (Nicastro et al., 2005). While we cannot exclude that there might be conditions in which a close conformation is stabilised (e.g. in the context of the full-length protein), the presence of a groove on the Josephin surface somewhat in proximity of the active site strongly suggests that this region could be involved in the recognition of protein substrates and/or of other Josephin partners, according to what already suggested in Nicastro et al. (2005). This hypothesis was inspired by homology with the mode of interaction observed in the structure of another ubiquitin cysteine protease, YUH1, in complex with ubiquitin-aldheyde (Johnston et al., 1999). In this structure, the substrate is accommodated in a groove formed by a helical hairpin spatially equivalent to that observed in Josephin. Experimental support was also directly provided by Mao et al. (2005), who mapped the interaction with ubiquitin by chemical shift perturbation onto the surface which includes the hairpin, although no attention was paid to the exact mode of binding and its consequences to the Josephin structure.

In conclusion, we believe that the 1yzb structure can be used as a reference for further studies of the Josephin domain and that the techniques presented here could be used routinely considered as a reference to validate the structural information available.

Acknowledgements

This work is supported by the EUROSCA European consortium.

References

Bax, A. (2003) *Protein Sci.*, **12**, 1–16.

Boulin, C.J., Kempf, R., Gabriel, A. and Koch, M.H.J. (1988) *Nucl. Instrum. Meth. A*, **269**, 312–320.

Brunger, A.T., Adams, P.D., Clore, G.M., DeLano, W.L., Gros, P. and Grosse-unstleve, R.W. et al. (1998) *Acta Crystallog. Sect. D*, **54**, 905–921.

Burnett, B., Li, F. and Pittman, R.N. (2003) *Hum. Mol. Genet.*, **12**, 3195–3205.

Chai, Y., Berke, S.S., Cohen, R.E. and Paulson, H.L. (2004) *J. Biol. Chem.*, **279**, 3605–3611.

Donaldson, K.M., Li, W., Ching, K.A., Batalov, S., Tsai, C.C. and Joazeiro, C.A. (2003) *Proc. Natl. Acad. Sci. USA*, **100**, 8892–8897.

Gabriel, A. and Dauvergne, F. (1982) *Nucl. Instrum. Meth.*, **201**, 223–224.

Habeck, M., Nilges, M. and Rieping, W. (2005a) *Phys. Rev. E. Stat. Nonlin. Soft Matter Phys.*, **72**, 031912.

Habeck, M., Nilges, M. and Rieping, W. (2005b) *Phys. Rev. Lett.*, **94**, 018105.

Habeck, M., Rieping, W. and Nilges, M. (2006) *Proc. Natl. Acad. Sci. USA*, **103**, 1756–1761.

Johnston, S.C., Riddle, S.M., Cohen, R.E. and Hill, C.P. (1999) *EMBO J.*, **18**, 3877–3887.

Kawaguchi, Y., Okamoto, T., Taniwaki, M., Aizawa, M., Inoue, M., Katayama, S., Kawakami, H., Nakamura, S., Nishimura, M. and Akiguchi, I. et al. (1994) *Nat. Genet.*, **8**, 221–228.

Koch, M.H.J. and Bordas, J. (1983) *Nucl. Instrum. Methods*, **208**, 461–469.

Konarev, P.V., Volkov, V.V., Sokolova, A.V., Koch, M.H.J. and Svergun, D.I. (2003) *J. Appl. Crystallogr.*, **36**, 1277–1282.

Kozin, M.B. and Svergun, D.I. (2001) *J. Appl. Crystallogr.*, **34**, 33–41.

Laskowski, R.A., Rullmannn, J.A., MacArthur, M.W., Kaptein, R. and Thornton, J.M. (1996) *J. Biomol. NMR*, **8**, 477–486.

Masino, L, Nicastro, G, Menon, RP, Dal Piaz, F, Calder, L. and Pastore, A. (2004) *J. Mol Biol.*, **344**, 1021–1035.

Masino, L., Musi, V., Menon, R.P., Fusi, P., Kelly, G., Frenkiel, T.A., Trottier, Y. and Pastore, A. (2003) *FEBS Lett.*, **549**, 21–25.

Mao, Y., Senic-Matuglia, F, Di Fiore, P.P., Polo, S., Hodsdon, M.E. and De Camilli, P. (2005) *Proc. Natl. Acad. Sci. USA*, **102**, 12700–12705.

Munishkina, L.A., Cooper, E.M., Uversky, V.N. and Fink, A.L. (2004) *J. Mol. Recognit.*, **17**, 456–464.

Nicastro, G., Menon, R.P., Masino, L, Knowles, P.P., McDonald, N.Q. and Pastore, A. (2005) *Proc. Natl. Acad. Sci. USA*, **102**, 10493–10498.

Nicastro, G., Masino, L., Frenkiel, T.A., Kelly, G., McCormick, J., Menon, R.P. and Pastore, A. (2004) *J. Biomol. NMR*, **30**, 457–458.

Porod, G. (1982) In *Small-angle X-ray scattering*, Glatter, O. and Kratky, O. (Eds.), Academic Press, London, pp. 17–51.

Rieping, W., Habeck, M. and Nilges, M. (2005a) *Science.*, **(309)**, 303–306.

Rieping, W., Habeck, M. and Nilges, M. (2005b) *J. Am. Chem. Soc.*, **127**, 16026–16027.

Sass, H.J., Musco, G., Stahl, S.J., Wingfield, P.T. and Grzesiek, S. (2000) *J. Biomol. NMR.*, **18**, 303–309.

Svergun, D.I. (1992) *J. Appl. Crystallogr.*, **25**, 495–503.

Svergun, D.I. (1999) *Biophys. J.*, **76**, 2879–2886.

Svergun, D.I. and Koch, M.H.J. (2003) *Rep. Progr. Phys.*, **66**, 1735–1782.

Svergun, D.I., Barberato, C. and Koch, M.H.J. (1995) *J. Appl. Crystallogr.*, **28**, 768–773.

Svergun, D.I., Petoukhov, M.V. and Koch, M.H.J. (2001) *Biophys. J.*, **80**, 2946–2953.

Taylor, J.P., Hardy, J. and Fischbeck, K.H. (2002) *Science*, **296**, 1991–1995.

Vestergaard, B., Sanyal, S., Roessle, M., Mora, L., Buckingham, R.H., Kastrup, J.S., Gajhede, M., Svergun, D.I. and Ehrenberg, M. (2005) *Mol. Cell*, **20**, 929–938.

Volkov, V.V. and Svergun, D.I. (2003) *J. Appl. Crystallogr.*, **36**, 860–864.

Vriend, G. (1990) *J. Mol. Graph.*, **8**, 52–56.