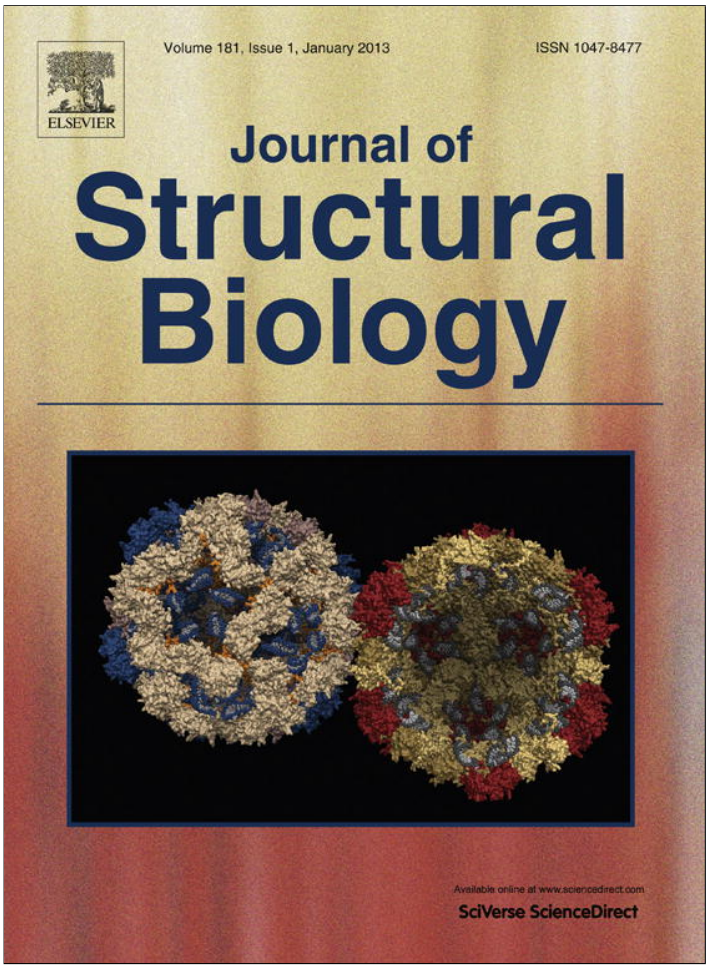
An assignment of intrinsically disordered regions of proteins based on NMR structures

ARTICLE *in* JOURNAL OF STRUCTURAL BIOLOGY · NOVEMBER 2012

Impact Factor: 3.23 · DOI: 10.1016/j.jsb.2012.10.017 · Source: PubMed



Journal of Structural Biology 181 (2013) 29–36

Contents lists available at SciVerse ScienceDirect

Journal of Structural Biology

journal homepage: www.elsevier.com/locate/yjsbi

An assignment of intrinsically disordered regions of proteins based on NMR structures

Motonori Ota [a,*], Ryotaro Koike [a], Takayuki Amemiya [a], Takeshi Tenno [b], Pedro R. Romero [c], Hidekazu Hiroaki [b], A. Keith Dunker [c], Satoshi Fukuchi [d]

[a] Graduate School of Information Sciences, Nagoya University, Nagoya 464-8601, Japan
[b] Graduate School of Pharmaceutical Science, Nagoya University, Nagoya 464-8601, Japan
[c] Center for Computational Biology and Bioinformatics, Indiana University School of Medicine, Indianapolis, IN 46202, USA
[d] Faculty of Engineering, Maebashi Institute of Technology, Maebashi 371-0816, Japan

ARTICLE INFO

Article history:
Received 3 July 2012
Received in revised form 26 October 2012
Accepted 30 October 2012
Available online 7 November 2012

Keywords:
Intrinsically disordered proteins
Deviations
Missing residues
Matthews's correlation coefficient
Nuclear Magnetic Resonance

ABSTRACT

Intrinsically disordered proteins (IDPs) do not adopt stable three-dimensional structures in physiological conditions, yet these proteins play crucial roles in biological phenomena. In most cases, intrinsic disorder manifests itself in segments or domains of an IDP, called intrinsically disordered regions (IDRs), but fully disordered IDPs also exist. Although IDRs can be detected as missing residues in protein structures determined by X-ray crystallography, no protocol has been developed to identify IDRs from structures obtained by Nuclear Magnetic Resonance (NMR). Here, we propose a computational method to assign IDRs based on NMR structures. We compared missing residues of X-ray structures with residue-wise deviations of NMR structures for identical proteins, and derived a threshold deviation that gives the best correlation of ordered and disordered regions of both structures. The obtained threshold of 3.2 Å was applied to proteins whose structures were only determined by NMR, and the resulting IDRs were analyzed and compared to those of X-ray structures with no NMR counterpart in terms of sequence length, IDR fraction, protein function, cellular location, and amino acid composition, all of which suggest distinct characteristics. The structural knowledge of IDPs is still inadequate compared with that of structured proteins. Our method can collect and utilize IDRs from structures determined by NMR, potentially enhancing the understanding of IDPs.

1. Introduction

Intrinsically disordered or natively unstructured proteins (IDPs) (Wright and Dyson, 1999) have been drawing considerable attention during the past decade (Dunker et al., 2002, 2001; Dyson and Wright, 2005; Tompa, 2005; Uversky and Dunker, 2010). These proteins do not adopt unique and specific three-dimensional (3D) structure under physiological conditions, especially when isolated from binding partners. IDPs are roughly classified, depending on the amount of intrinsically disordered regions (IDRs) they contain, into fully disordered and partially disordered proteins. Some segments of IDRs can become structured via interaction with their binding partners (Fukuchi et al., 2012; Fuxreiter et al., 2004; Galea et al., 2008; Gould et al., 2010; Mohan et al., 2006) in a process called coupled folding and binding (Wright and Dyson, 1999). Other IDRs remain flexible, and sometimes act as linkers between structural domains (Tompa, 2005). IDPs are more abundant in eukaryotes than in prokaryotes, and eukaryotic IDPs are localized preferentially in the nucleus rather than the cytoplasm of cells (Fukuchi et al., 2011; Minezaki et al., 2006; Ward et al., 2004). IDPs are frequently involved in transcription, translation, signal transduction, and phosphorylation (Dunker et al., 2002; Dyson and Wright, 2005; Iakoucheva et al., 2002; Tompa, 2005), and many are associated with disease (Cheng et al., 2006). These observations indicate that IDPs are crucial for complex biological phenomena exhibited by higher-order organisms. Because of this importance, much effort has been devoted, both in experimental and theoretical research, to reveal the structural characteristics and functional nature of IDPs.

When one performs a computational study on IDPs, suitable datasets of IDPs are essential (Fukuchi et al., 2012; Sickmeier et al., 2007). Conventionally, IDPs and IDRs are collected one-by-one from literature searches (Sickmeier et al., 2007) or, more efficiently, by identifying so-called missing residues in X-ray

Abbreviations: IDP, intrinsically disordered protein; IDR, intrinsically disordered region; PDB, Protein Data Bank; NMR, Nuclear Magnetic Resonance; MCC, Matthews's correlation coefficient.

* Corresponding author. Fax: +81 52 789 4782.
E-mail address: mota@is.nagoya-u.ac.jp (M. Ota).

1047-8477/$ - see front matter © 2012 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.jsb.2012.10.017

crystalline structures of proteins stored in the Protein Data Bank (PDB) (Berman et al., 2003). Local static or dynamic disorder commonly leads to missing residues. Because of the highly flexible nature of IDPs, crystallizations of IDPs or proteins with long IDRs are difficult in general. Thus, IDRs determined by use of missing residues are usually limited to somewhat short IDRs flanked by crystallizable structural domains. On the other hand, information on very long IDRs, including fully disordered proteins, can be collected mostly by reading experimental studies in the literature (Sickmeier et al., 2007; Uversky et al., 2000). Obviously, this procedure is time-consuming and laborious, as compared to the collection of missing-residues, which is easily performed by parsing PDB files. The collected data provide fundamental information for various studies (Fukuchi et al., 2012; Sickmeier et al., 2007), e.g., sequential, structural, functional, and evolutionary analyses, and for the development of sequence-based disorder predictors (He et al., 2009).

Although missing residues in X-ray structures are effectively employed in computational studies of IDPs (Lobanov et al., 2010), efficient methods to gather IDR information from Nuclear Magnetic Resonance (NMR) data have not yet been developed. Using NMR-based disorder information is essential for a better understanding of IDPs, because of the following points: First, the amount of protein structure determined by NMR is not negligible. The number of NMR entries in the PDB is greater than 8000, and this is more than 10% of all PDB entries (PDB statistics are available at http://www.rcsb.org/pdb/statistics/holdings.do) (Berman et al., 2003). Second, NMR can potentially be used to study the dynamical nature of long IDRs including wholly disordered proteins (Mittag and Forman-Kay, 2007), whereas X-ray crystallography involves less fluctuating structures, and thus IDRs found in X-ray structures tend to be limited to relatively short segments. Conventional methods based on $^{1}H–^{15}N$ heteronuclear nuclear overhauser effects (NOEs) have been used to identify IDRs in proteins. Recently, a new methodology of NMR, relaxation dispersion spectroscopy, has been applied to the study of IDPs' interactions and the coupled folding and binding process, a characteristic feature of IDPs (Sugase et al., 2007a,b). Finally, the proteins whose structures are only analyzed by NMR may contain many IDPs because structures determined only by X-ray crystallography and only by NMR are biased towards cytoplasmic and nuclear proteins, respectively (see Fig. S1 in Supplementary data). As nuclear proteins contain many IDPs (Fukuchi et al., 2011; Minezaki et al., 2006; Ward et al., 2004), ignoring NMR data could result in neglecting a substantial fraction of structural information on IDPs. All these observations indicate that structural data provided by NMR are crucial to the study of IDPs. A suitable approach is thus necessary to locate IDRs and IDPs based on NMR data, so that results can be integrated with disorder information derived from X-ray structures.

One may note that proteins with $^{1}H–^{15}N$ NOE data have been collected in the Biological Magnetic Resonance Bank (BMRB) (Ulrich et al., 2008), but unfortunately, the number of entries containing $^{1}H–^{15}N$ NOE data is only 173 (as of June, 2012), significantly smaller than the NMR structures in the PDB (more than 8000). Cleary, it is reasonable to focus on the entries in the PDB, in terms of data size quantity, rather than information based on the $^{1}H–^{15}N$ NOE spectra.

The goal of the present study is to develop a systematic method to locate IDRs using NMR structures in the PDB, as this would provide a new, large source for further study. Normally, a PDB entry produced from NMR contains a number of models for an identical protein. By examining multiple model structures, we can locate residues with coordinates that show high deviation in the various models. Such coordinate deviations arise because the data emanating from those regions is very sparse and so can be fit by multiple models. Such lack of data can arise because the given region is indeed intrinsically disordered and thus provides sparse data. Alternatively, such lack of data could arise from a region of structure that is rigid but highly exposed to solvent, meaning that the region lacks nearby protons to provide the spin–spin interactions that give rise to the NMR information used to determine precise protein structure.

For this initial study, we assume that regions of larger deviation in NMR studies are most often indicators of IDRs. Based on this assumption, we first determined the single deviation threshold as the first-order approximation that produces the highest correlation between IDRs located from both NMR and X-ray structures of the same protein, regardless of experimental conditions. We recognized structural ensembles depend on structure refinement protocols and software, but these characteristics were not taken into account for this study. Note that several other researchers have compared NMR and X-ray structures, but these previous comparisons focused on information regarding the structural aspects of proteins (Brunger, 1997; Garbuzynskiy et al., 2005; Serrano et al., 2010; Sikic et al., 2010), whereas here the focus is on the IDRs. Our approach unveils putative IDRs in NMR structures in a way that can be fully automated.

Given these new IDR data, we then compared characteristic features of IDRs derived from NMR structures to those of IDRs derived from X-ray structures, in terms of sequence length, IDR fraction, protein function, cellular location, amino acid composition and post-translational modifications. Finally, for further validation, we carried out additional comparisons with $^{1}H–^{15}N$ NOE data collected mainly from literature, which gives an identification of IDRs. Overall, the IDR data revealed from the automated analysis of NMR structures described herein helps to broaden and improve our understanding of proteins lacking stable structure.

2. Materials and methods

2.1. Missing residues

In PDB entries, protein sequences used in the structure determination are declared in the SEQRES lines, whereas the sequence of residues whose coordinates could be located can be found in the ATOM lines. By aligning the sequences obtained from these two sources, we detected the missing residues as those found in the gap sites of the alignment (Jones and Ward, 2003). This same information also can be found in convenient form on the web (http://dunbrack.fccc.edu/xml2pdb.php).

2.2. Deviation of residues (RMSD)

We evaluated the deviations of residues from PDB entries of NMR structures containing multiple models. The root mean squared deviation (RMSD) of residue at the position j (Δr_j) is defined using the coordinates of Cα atoms from a set of model structures, as,

$$\Delta r_j = \sqrt{\sum_{i}^{3}\sum_{m}^{M}\frac{\left(x_{i,j}^{m}-\bar{x}_{i,j}\right)^2}{M}} \quad (1)$$

where, $x_{i,j}^{m}$ is the i-th coordinate value of Cα atom in the residue at position j in the model m, M is the total number of models, and $\bar{x}_{i,j}$ is the average of i-th coordinate value of Cα atom at j over all M models. We relied on the superimposed coordinates in the selected PDB entries (see Dataset section in detail).

2.3. Dataset

From the PDB (Berman et al., 2003), we selected proteins whose structures have been determined by X-ray crystallography and/or NMR, mapping their sequences onto the corresponding SwissProt sequences (The UniProt Consortium, 2011), and removing artificial

sequence tags. We selected X-ray structures containing at least 30 amino acid residues for which 3D coordinates are available at higher than 3.0 Å resolution, and at least 5 amino acid residues for which the coordinates are unavailable (missing residues, see Section 2.1). We used only monomeric proteins, as judged by protein quaternary structure (PQS) (Henrick and Thornton, 1998). We selected the PDB entries of NMR structures including only a single protein chain (assumed to be monomeric), and multiple models longer than at least 30 amino acid residues. We compared the protein sequences thus selected from X-ray and NMR structures using BLAST (Altschul et al., 1997), and paired them whenever a pair of proteins met the following conditions: the alignment covered 90% or more of the total length and the sequence identity was greater than 95%; both structures of a protein pair should either contain ligand molecules (both holo form), or nothing (both apo form); and the author names should be at least partially different. We set the last condition in order to avoid the selection of structural models determined in the same study, because their structural features, e.g., IDRs, are frequently identical. Based on the selected protein pairs, we performed single-linkage clustering, discarding singletons. We regarded one cluster as the set of multiple structures determined by X-ray crystallography and/or NMR for a single protein. We partitioned the protein-pair set into three categories: (1) proteins whose structures were multiply determined only by X-ray crystallography (the X-ray dataset); (2) proteins whose structures were multiply determined only by NMR (the NMR dataset); and (3) proteins whose structures were determined by both X-ray crystallography and NMR (the both-ways dataset). We discarded extremely flexible NMR structures, where the minimum deviation (RMSD) of residues in the structure was more than 6 Å, as well as very rigid NMR structures, where the maximum RMSD was less than 3 Å, so that only clear deviations were considered.

2.4. Correlation of intrinsically disordered regions (IDRs)

In the X-ray structures, we regarded residues missing from the coordinate data as belonging to IDRs (D state), and residues with coordinates as ordered (O state). For each residue in the NMR structures, we calculated its deviation as RMSD, and if the deviation was larger than a set threshold, Δr_{th}, the residue was labeled as disordered, or part of an IDR (D), and as ordered (O) otherwise. We also labeled the missing residues in NMR structures as disordered (D). We determined the best value for the threshold Δr_{th} as described below. When we prepared two structures of a protein, we labeled the state (O/D) on each residue, regardless of the method of structure determination. When the state of a residue in one structure is A, and that of another is B, then we denote the state of the residue is (A, B), where A and B are one of O or D. We evaluated the agreement of the residues' states in a protein pair by employing the Matthews's correlation coefficient (MCC) (Matthews, 1975), defined as,

$$\mathrm{MCC} = \frac{N(\mathrm{O},\mathrm{O}) \times N(\mathrm{D},\mathrm{D}) - N(\mathrm{O},\mathrm{D}) \times N(\mathrm{D},\mathrm{O})}{\sqrt{N(\mathrm{O},^{*}) \times N(^{*},\mathrm{O}) \times N(\mathrm{D},^{*}) \times N(^{*},\mathrm{D})}} \quad (2)$$

where N(A, B) is the number of residues in state (A, B) in a protein, and $N(\mathrm{A},*) = N(\mathrm{A},\mathrm{A}) + N(\mathrm{A},\mathrm{B})$.

We compared the X-ray structures and NMR structures of identical proteins by applying the Eq. (2), and determined the best RMSD threshold, Δr_{th}, as the one that produced the highest MCC of IDRs between members of protein pairs.

2.5. Chou–Fasman parameter of IDRs

The amino acid compositions of IDRs are represented as the Chou–Fasman parameters (Chou and Fasman, 1978),

$$\mathrm{CF}(a) = \frac{N^{a}(\mathrm{D})/N^{a}_{\mathrm{all}}}{N(\mathrm{D})/N_{\mathrm{all}}} \quad (3)$$

where N^{a}(D) is the number of amino-acid residue a in IDRs, N^{a}_{all} is the total number of amino-acid residue a in the entire sequence, N(D) is the total number of residues in IDRs, and N_{all} is the total number of residues in the entire sequence.

2.6. Assessment of $^{1}H–^{15}N$ NOE peaks

For 24 proteins, we assessed the IDRs derived from NMR structures by comparison with their $^{1}H–^{15}N$ NOE data. We identified IDRs in NMR structures (from the NMR dataset plus the both-ways dataset) and the fluctuating residues of the corresponding proteins by means of $^{1}H–^{15}N$ NOE. We found only one entry of $^{1}H–^{15}N$ NOE in BMRB (Ulrich et al., 2008) for this purpose. The other data were taken from the literature manually. We considered $^{1}H–^{15}N$ NOE smaller than 0.5 (including negative values) indicated dynamically fluctuating residues. Our definition meant that the identified residues were fluctuating more rapidly, or exhibiting larger amplitude of the backbone fluctuation, than those that with $^{1}H–^{15}N$ NOE < 0.6–0.7, which were excluded as the exceptionally tumbling residues in the established NMR protocols of relaxation analysis (Larsson et al., 2003; Pawley et al., 2001; Sprangers et al., 2000).

3. Results and discussions

3.1. A definition of IDRs on NMR structures

We selected 489, 75 and 55 proteins as the X-ray, the NMR, and the both-way datasets, respectively, and calculated missing residues and/or residue deviations for the structures in the both-ways dataset. The averaged maximum RMSD and the standard deviation for the NMR structures of the both-ways dataset were 10.2 and 7.4 Å, respectively. We assumed a provisional threshold for deviation (Eq. (1)), and used it to label IDRs of NMR structures as explained in the methods section. We calculated the MCC (Matthews, 1975) for any given pair of X-ray and NMR structures (Eq. (2)). When multiple X-ray or NMR structures were available for a given protein, we obtained many MCCs due to the multiple pairings of X-ray and NMR structures. In this case, we considered as representative only the structure pair that gives the maximal MCC. By changing the threshold value in 0.1 Å increments, we searched for the threshold that produced the maximum MCC of O/D states in the entire set of X-ray and NMR structures.

As is apparent from Eq. (2), the denominator will be zero (some of the four Ns will be zero), if we chose an inappropriate threshold value: for instance, when we choose a very large threshold, all residues in NMR structure will be labeled as being ordered, and thus $N(*, \mathrm{D})$ will be zero. In the 55 proteins of the both-ways dataset, we could calculate all of the MCCs (i.e., denominators were non-zero), when we adjusted the threshold value within 1.1–3.3 Å. We also found that threshold values between 2.9 and 3.3 Å gave almost maximum MCC (greater than 0.62) when using their own set of representative structural pairs (Note that the set of representative X-ray and NMR structure pairs changes when we alter the threshold, because they depend on the threshold value). We fixed these 5 sets of representative structural pairs (determined at 2.9, 3.0, 3.1, 3.2 and 3.3 Å) to test each threshold, and averaged the results: Fig. 1a (circles and solid line) indicates that the average MCC at 3.2 Å is the highest (0.626). Thus, we set 3.2 Å as the RMSD threshold (Δr_{th}) used to label ordered and disordered residues in NMR structures. The representative structural pairs determined at 3.2 Å (Table S1 in Supplementary data) were employed to derive MCC (triangles and dashed line in Fig. 1a), and also the ROC plot

(Fig. 1b). The cross in Fig. 1b indicates the result at the 3.2 Å threshold. The true positive rate (the D state of NMR structure coincides with the D state of X-ray structure) is 0.628, and the false positive rate (the D state of NMR structure is falsely assigned to the O state of X-ray structure) is 0.023. Variation between the X-ray and NMR structures is discussed in the following paragraphs with examples. To examine the dependency on the number of multiple models, we selected 21 PDB entries containing 20 NMR models (Table S1), employed only first 10 models, and compared the results with our original ones (using 20 models). The assignment of IDR was mismatched only for 0.6% of total regions (18/2716), and the average difference of RMSD is 0.09 Å. This study suggests that the assignments are robust regardless of the number of models (10 or 20). Note that to calculate MCC in Fig. 1, we combined the 55 proteins in the set to make one large virtual sequence, to which we applied Eq. (2). The individual results of each of the 55 proteins are summarized in Table S1 in Supplementary data.

To demonstrate the relationship between the missing residues and the highly deviating residues, we present a couple of examples. Fig. 1c shows the structures of translationally controlled tumor protein (TCTP), determined by X-ray crystallography (left, PDB ID: 1yz1A) and NMR (right, 2hr9A (Feng et al., 2007)). In the X-ray structure, the coordinates for residues 39–66 are missing, whereas in the NMR structure, more than 3.2 Å deviations are observed for residues 41–64 (Fig. 1e). The MCC for the order and disordered regions of the structures is 0.913, indicating that the missing residues and highly deviating residues largely coincide, even though the IDR is located not at the termini but in the middle of the sequence. By contrast, in the case of chemosensory protein Csp2 (Fig. 1d), the obtained MCC was low (0.354). In the X-ray structure (left, PDB ID 1kx8A (Lartigue et al., 2002)), the coordinates for both terminal regions (1–9, and 110–112) are missing (Fig. 1f). In the NMR structure (right, PDB ID 1k19A (Mosbah et al., 2003)), the N-terminal residues (1–6) are highly deviating,

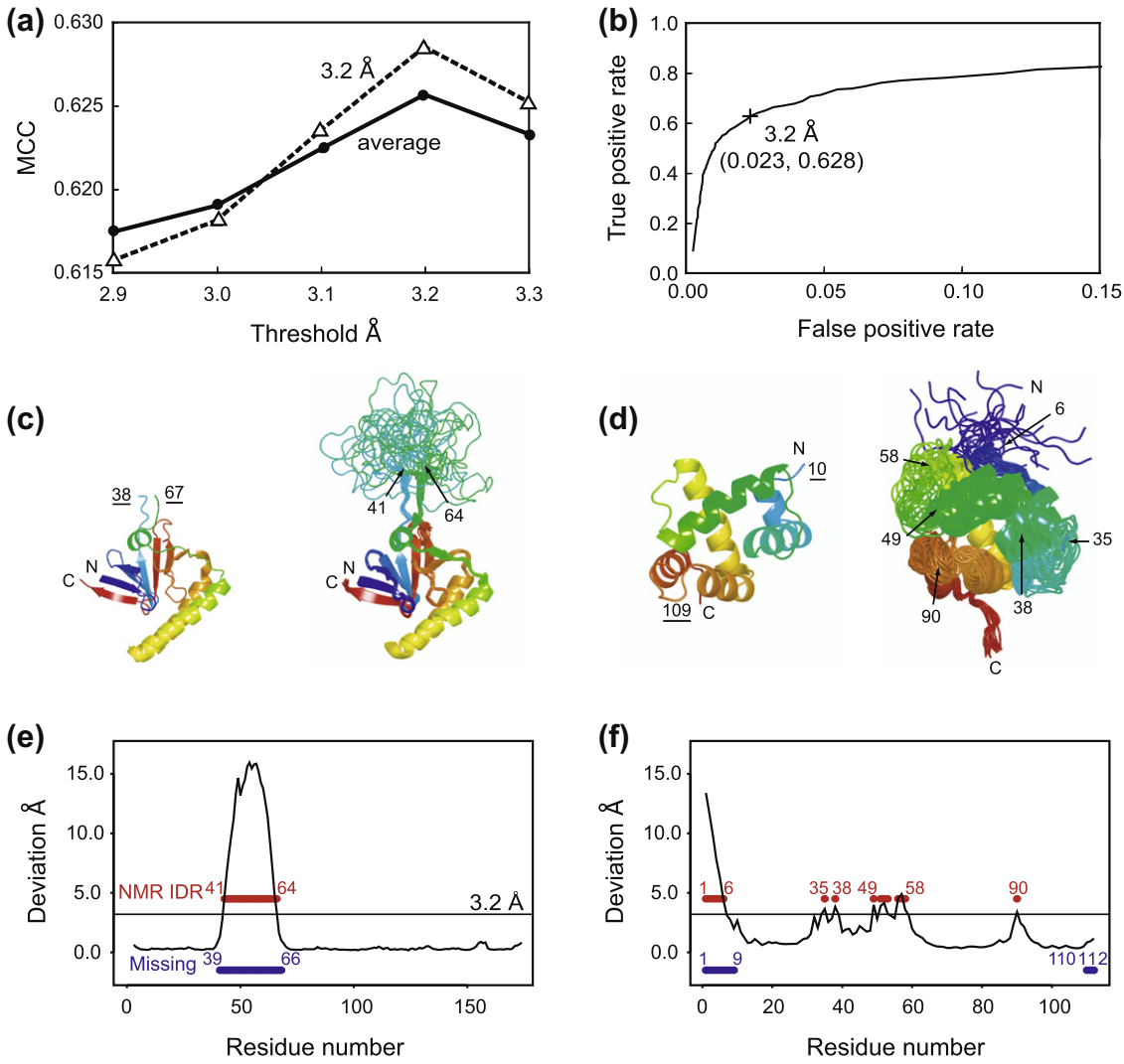


Fig. 1. Determination of deviating residues in NMR structure that correspond to missing residues in X-ray structure. (a) MCC for order and disorder regions in X-ray and NMR structures calculated with different threshold values employing a set of representative pairs at 3.2 Å (triangles and dashed line) and 5 representative sets (circles and solid line, see manuscript for details). (b) ROC plot derived using representative structures at 3.2 Å threshold value. Plus symbol indicates the position at the best threshold. (c) X-ray (left, 1yz1A) and NMR (right, 2hr9A (Feng et al., 2007)) structures of translationally controlled tumor protein (TCTP). The terminal residues of ordered regions are marked with the underlined numbers. The start and end residues of disordered regions are marked with the normal numbers. (d) X-ray (left, 1kx8A (Lartigue et al., 2002)) and NMR (right, 1k19A (Mosbah et al., 2003)) structures of chemosensory protein Csp2. Ordered and disordered regions are shown in the same manner as (c). (e) The RMSD of NMR structures of TCTP. The 3.2 Å threshold is shown as a horizontal line. The NMR assigned disordered regions are shown as red lines, and the missing residues of the X-ray structure are shown as blue lines. (f) The deviations of chemosensory protein Csp2 shown in the same manner to (e).

which agreed with the X-ray structure, but the C-terminal residues are not. In addition, some residues mainly at the flanking loops of the third helix (38–53) are highly deviating (Fig. 1d and f). These regions are flexible so that the third helix moves and exposes the hydrophobic core for lipidic ligand binding (Mosbah et al., 2003). From visual inspection, we also suspect that crystal contacts are likely to stabilize these flexible regions in the X-ray structure. In fact, the stabilization of these regions by crystal contacts was reported for this protein's other PDB entry (1kx9 (Lartigue et al., 2002)), but the packing patterns of molecules in 1kx8 and 1kx9 are different.

In general, a greater source of variation between the X-ray and NMR structures is likely to be the fact that IDRs are sensitive to the crystallization conditions: indeed, it is often observed that a particular region is structured when a protein is crystallized under one set of conditions and disordered under another set of conditions (Le Gall et al., 2007; Zhang et al., 2007; Mohan et al., 2009). This sensitivity to crystallization conditions means that some segments identified as disordered in NMR experiments become structured as a result of crystallization. Identifying such regions will be a goal of future studies.

3.2. IDRs in NMR structures

The addition of NMR structures to the study of IDP examples would potentially provide new insights if IDRs found mostly in NMR structures have distinct characteristics as compared to those prevalent in X-ray structures. To study whether this is the case, we examined IDPs from both NMR and X-ray structures in terms of length, function, and amino acid composition.

To define reliable IDRs, we selected multiple PDB structures for a given protein, and collected results as follows: we identified highly deviating residues (>3.2 Å) and missing residues for 75 proteins in the NMR dataset. Since at least two PDB structures were determined for any given protein in the NMR dataset by different research groups (see the Dataset section), the IDRs obtained can vary slightly among these structures. Thus, we selected the structural pair that shows the highest MCC of IDRs for each protein (Eq. (2)), and regarded the pair as the representative of the protein. Highly deviating or missing residues in any of the structures in the representative pair were considered as IDRs. As two structures (a pair) were taken for each of 75 proteins, a total of 150 NMR structures were considered (the "representative NMR" (rNMR) dataset). To ensure that our NMR IDR dataset contains substantial ordered and disordered regions, we limited structural variation as follows: we discarded NMR structures which did not contain at least 10 successive ordered residues or whose longest ordered region was shorter than 10% of the total sequence length; also, when an IDR contains more than one secondary structural element, we eliminated the structure, because structural superposition of models in the PDB may be unsuitable to judge the deviation (e.g., calmodulin, PDB ID: 1dmoA (Zhang et al., 1995)).

We identified missing residues for the structures of 489 proteins in the X-ray dataset. We selected the structural pair that shows the highest MCC of missing residues (Eq. (2)) for each protein, and regarded their missing residues as IDRs and the structure pair as representative. In total, we considered 978 X-ray structures (the "representative X-ray" (rX-ray) dataset).

3.2.1. Length of proteins and IDR portions

The total lengths of PDB entries are plotted in Fig. 2a for the rNMR and the rX-ray datasets. The average lengths are 96 and 353, respectively, for the rNMR and the rX-ray dataset. 88% of the NMR structures are shorter than 150 residues. The shorter length for NMR-determined as compared to X-ray-determined structures was fully expected. That is, determination of small

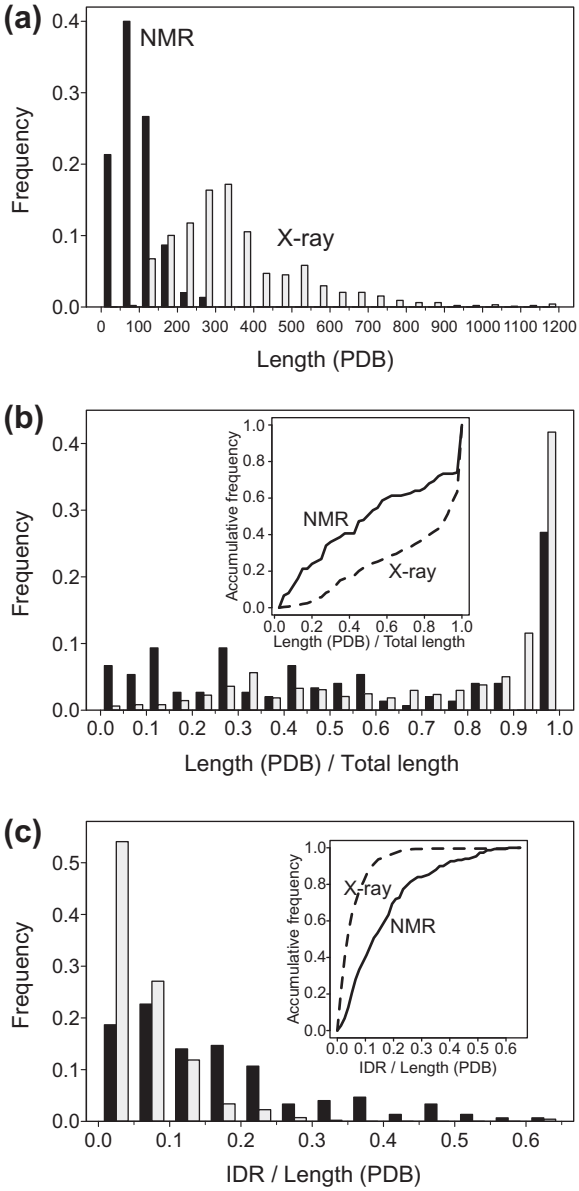


Fig. 2. Features of X-ray and NMR structures examined in this study, in terms of sequence length. (a) Sequence length of proteins deposited in the PDB. Gray and black bars correspond to the X-ray and NMR structures, respectively. (b) Ratio of sequence lengths used in the PDB to the total length of proteins. Inset indicates the cumulative frequencies. (c) Ratio of IDR lengths to the sequence lengths in the PDB.

proteins by NMR, say less than 25 kDa, is well established, but determination of larger protein structures remains very challenging due to the difficulty of assigning so many NMR peaks and other technical problems (Tzakos et al., 2006).

One of the factors contributing to the small size of the NMR structures is that many of the NMR structures were determined for only a portion of long multi-domain proteins. Fig. 2b shows the ratio of the sequence used for the structural determination (sequence length on SEQRES lines in PDB entries) to the total protein length. About half structures in the rNMR dataset were determined for segments shorter than half the full sequence of the proteins (see the cumulative fractions in the inset). The ratios of IDR lengths to total sequence length for our representative datasets are plotted in Fig. 2c. Even though the definition of IDRs in NMR structures was adjusted to be similar to that of X-ray structures, NMR structures have a higher content of IDRs than X-ray structures, reflecting

the flexible nature of proteins in the rNMR dataset. It may be also due to the missing restraint in NMR structure, or crystal-packing artifact in X-ray structure (see Fig. 1d). We noticed that IDRs in rNMR dataset were abundant at the sequence termini, and it was partly because these IDRs were likely to be domain-linkers of multi-domain proteins. But, even though we ignore the samples that only have IDRs at termini, there are also a certain number of IDRs at the middle of sequence (Fig. S2 in Supplementary data). On the other hand, proteins in the rX-ray dataset tend to be rigid, and contain only a small portion of IDRs. This observation supports the conjecture that the requirement for crystallization selects against proteins having larger amounts of disorder (Le Gall et al., 2007).

3.2.2. Function of IDPs in rNMR dataset

Relying on the Gene Ontology (GO) functional annotations in UniProt (The UniProt Consortium, 2011), we characterized the 75 IDPs in the rNMR dataset. The GO terms are grouped into three categories, location, biological process, and molecular function. We counted the number of occurrence for each GO term in the datasets, and normalized them by the total number of occurrence in each of the categories. The frequencies of GO terms were compared with that of the rX-ray dataset (489 proteins). Fig. 3 (black bars) depicts the terms showing significant frequency differences for each category between datasets. In the location category (Fig. 3a), "nucleus" was enriched in IDPs of the rNMR dataset. This suggests a preference of NMR structural analyses to nuclear proteins. In contrast, "cytoplasm" was significantly enriched in the rX-ray dataset. In the category of biological process (Fig. 3b), "transcription", "transcription regulation" and "DNA replication" were enriched in the rNMR dataset. These terms are well associated with nuclear proteins, and thus these results agree with those from the location category. In the category of molecular function (Fig. 3c), IDPs in the rNMR dataset were more related to "prion" and "hormone", implying the association of the IDPs with diseases. On the other hand, "hydrolase", "transferase", and "kinase" were frequently found in the rX-ray dataset, meaning that enzymes are more suitable for analysis by X-ray crystallography. To investigate the features of IDPs analyzed by NMR in more detail, we prepared 104 proteins that satisfied the conditions in Dataset section as rNMR dataset did, but contained at most 4 disordered residues (thus they are non-IDP). We compared GO terms of them to that of rNMR dataset. "Nucleus", "transcription", "transcription regulation", "DNA replication", "prion" and "hormone" were also enriched in rNMR dataset (see gray bars in Fig. 3), showing that these enriched GO terms originate from not only the experimental methodology (NMR or X-ray), but also the nature of protein (IDP or not). On the other hand, "secreted" is significantly enriched in non-IDPs analyzed by NMR (see also Fig. S1).

3.2.3. Amino acid composition of IDRs

The amino acid compositions of IDRs in the rNMR and the rX-ray datasets were calculated independently, and represented by Chou–Fasman parameters (CFPs, Eq. (3)). A scatter plot of CFPs is shown in Fig. 4a. Both the rX-ray and the rNMR datasets indicate similar tendencies, i.e., the CFPs of hydrophobic amino acids are underrepresented and those of hydrophilic amino acids are overrepresented, and they are also essentially consistent with the order-promoting and disorder-promoting tendencies of amino acids (Dunker et al., 2001). Although the Pearson's correlation coefficient between the CFPs of the two datasets is 0.84, Fig. 4b, in which we subtracted CFPs of rX-ray dataset from those of rNMR dataset, shows some differences between the corresponding CFP values. In Fig. 4b, a positive value of a specific amino-acid residue indicates that the residue is more frequent in IDRs of the rNMR dataset, and negative value indicates the opposite. We found that the CFPs of the aromatic amino acids that were considered order-promoting

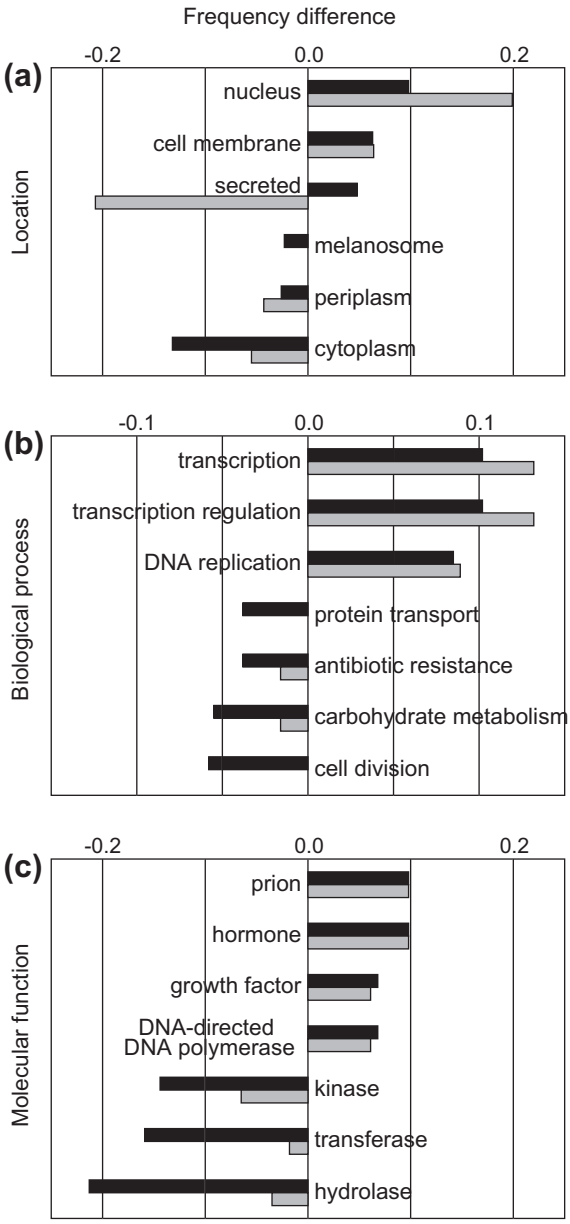


Fig. 3. Differences between proteins in the rX-ray and the rNMR datasets examined in this study, in terms of GO annotations (black bars). Annotated GOs were counted in each category (X-ray or NMR) and converted into the observed frequencies. Differences of frequencies are shown in the horizontal axis. Positive value indicates GO terms more abundant in NMR structures. (a) Location. (b) Biological process. (c) Molecular function. The same analyses were performed using rNMR datasets and ordered proteins investigated by NMR (gray bars, see the manuscript for details).

amino acids (Dunker et al., 2001), were larger in the rNMR dataset. In the statistics of Trp or Phe, each of them was less frequent in terms of amino-acid composition (N^W_{all}/N_{all} or N^F_{all}/N_{all} in Eq. (3)) than that of the rX-ray dataset, resulting in the larger CFPs. On the other hand, Tyr in the rNMR dataset was more frequent in both the entire sequence (N^Y_{all}/N_{all}) as well as the IDR ($N^Y(D)/N(D)$) than that of the rX-ray dataset. We noticed that some Tyr in IDRs served as phosphorylation sites. For example, Tyr137 of coactosin-like protein (PDB ID: 1wm4A (Hellman et al., 2004) and 1 udmA (Goroncy et al., 2009)), whose deviation is larger than 8.5 Å, is identified as a phosphorylation site through a high-throughput approach (Hornbeck et al., 2012). Some Ser and Thr in IDRs were also phosphorylation sites, for instance Ser32 and Ser34 of parathyroid hormone (1hpyA (Marx et al., 2000)). Lys is also enriched in the IDRs of the rNMR dataset, and some of them are sites for post-

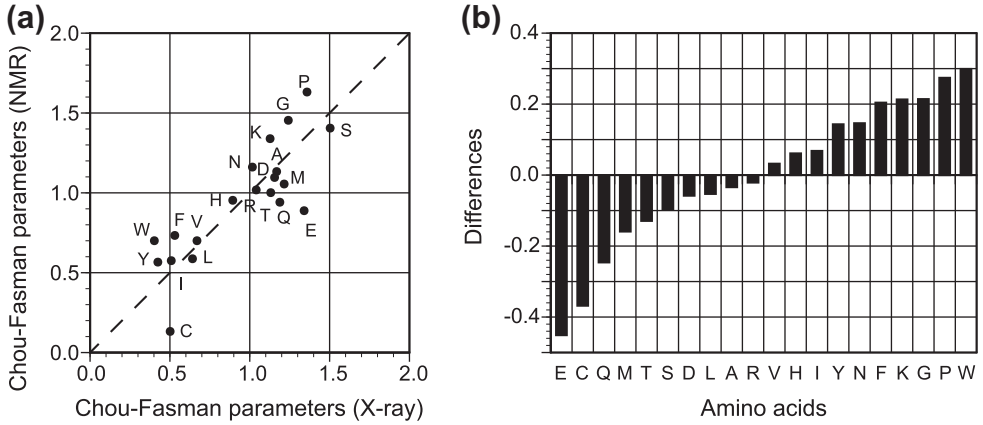


Fig. 4. Amino acid compositions of IDRs represented as Chou-Fasman parameters (CFPs) derived from either X-ray or NMR structures. (a) Scatter plot. (b) Differences. Positive value indicates CFPs of NMR structures are larger than those of X-ray structures.

translational modification. For example, a neddylation, was reported (Lee et al., 2008) on Lys699 of amyloid beta A4 protein (1z0qA (Tomaselli et al., 2006)), whose deviation is larger than 5.3 Å. Neddylation is the post-translational addition of a protein, NEDD8, that is closely related to ubiquitin, and so is similar to ubiquitination but with distinctive biological consequences (Rabut and Peter, 2008). On the other hand, the CFP for Glu in the rNMR dataset is smaller than that of the rX-ray dataset. We investigated this discrepancy, and found that the Glu within the rNMR dataset showed higher preference for location in helices (Chou and Fasman, 1978) and thus they tended to be ordered (data not shown).

The robustness of disordered regions to changes in sequence or environmental conditions is not completely known, but it has been shown (Mohan et al., 2009) that the structures and, correspondingly, the IDRs obtained from diverse X-ray experiments can vary greatly from one structure to the next depending on their experimental conditions and even small changes in sequence. This is, of course, true also of NMR experiments, where the pH and buffer conditions in solution might lead to disorder-to-order or order-to-disorder transitions as compared to the disorder content presented in any given X-ray structures deposited in the PDB.

3.3. Defined IDRs in NMR structure and experimental fluctuations

We examined how IDRs derived from NMR correspond to dynamically fluctuating residues in solution by means of $^{1}H–^{15}N$ NOE data (Muchmore et al., 1996). We gathered corresponding $^{1}H–^{15}N$ NOE experiments for 24 NMR structures from literature (the list of references, see Table S2 in Supplementary data). Note that experimental data for only one of them (neuropeptide Y) was deposited in the BMRB (as bmr548) (Ulrich et al., 2008). In the 2173 residues in 24 proteins, 400 residues were identified as IDRs by our definition (more than 3.2 Å RMSD). Of these residues, 320 (80%) had an $^{1}H–^{15}N$ NOE peak smaller than 0.5. The other 80 IDR residues, which displayed $^{1}H–^{15}N$ NOE values greater than 0.5, correspond mostly to local minima of the $^{1}H–^{15}N$ NOE profiles. We also noticed that around 80% of them are located within loop regions. A total of 1773 $^{1}H–^{15}N$ NOE peaks were collected for regions identified as being structured (less than 3.2 Å RMSD). Of these, just 130 (7%) have peak values smaller than 0.5, showing that the identification of structure by our method gives excellent agreement with assignment of structure by relaxation methods. Most of these (92 of 130) are within 3 residues from the termini of detected IDRs. For IDRs by our definition and $^{1}H–^{15}N$ NOE experiments, the MCC is 0.695.

For historical reasons, the NMR structure of a protein is commonly represented as a structural ensemble of ten to twenty conformations that equally satisfy the NMR-derived experimental data (Markley et al., 1998). Thus, as mentioned in the introduction, poor structural convergence is due to an insufficient number of structural constraints, and does not necessarily correspond to a protein's increased flexibility or intrinsic disorder in solution. In addition, as structural ensembles depend on refinement protocols and software etc., extensive validations are required for NMR structures (Doreleijers et al., 2012). Nevertheless, the $^{1}H–^{15}N$ NOE data described above suggest that a very large proportion of the residues identified by our method are highly fluctuating, thus indicating that structured regions yielding sparse data are rare in our dataset. This comparison between structural deviations with the $^{1}H–^{15}N$ NOE data suggests that our method for identifying IDRs is, for the most part, valid and reasonable, and complements the unavailability of $^{1}H–^{15}N$ NOE data.

4. Conclusions

Here we propose a systematic method for the assignment of IDRs based on NMR structures by calculating the deviation (as RMSD) of residues in multiple models. Based on the missing residues in corresponding X-ray structures, we determined a threshold of RMSD (3.2 Å) that exhibits the highest correlation between the missing residues in the X-ray structures and the identified IDRs in the NMR structures. We applied this definition to IDPs whose structures were only determined by NMR, and examined their features. We found they are likely to be nuclear proteins involved in transcription and regulation, or prion or hormone related proteins. Their NMR structures are likely to be small domains of large proteins, containing more IDRs as compared to X-ray structures. The NMR IDRs were enriched in aromatic residues, a part of which seems to correspond to post-translational modification sites.

Since the proposed 3.2 Å threshold value relies on the PDB dataset, we believe it should be reassessed when the number of proteins in the dataset increases in the future. However, we would like to emphasize that our procedure using multiple models in NMR structures is a reasonable way to identify a new collection of IDRs, as confirmed by the comparison with experimentally determined fluctuation in solution. Meanwhile, we still anticipate the potential of $^{1}H–^{15}N$ NOE, as it is the basic information for IDRs. Since December of 2010, chemical shift data should be deposited in BMRB at the submission of PDB. However, among 720 BMRB entries under the rule, only 20 contain $^{1}H–^{15}N$ NOE (Ulrich et al., 2008). Submission of $^{1}H–^{15}N$ NOE data into BMRB should be

strongly encouraged. In addition, comparison between identified IDRs by our method and the predictions of IDRs (He et al., 2009), or the predictions of random coils (Tamiola et al., 2012), could be informative and will be conducted in the near future.

Incorporation of new IDRs from NMR data will potentially expand the scope of our knowledge regarding IDPs and provide new insights to their study, as implied by the differences found in this study between NMR and X-ray disorder. For instance, we plan to apply our definition to both monomeric as well as oligomeric NMR structures, anticipating that the latter will include proteins presenting order/disorder transitions via the coupled folding and binding mechanism (protean segments, MoRFs, or ELMs) (Fukuchi et al., 2012; Fuxreiter et al., 2004; Galea et al., 2008; Gould et al., 2010; Mohan et al., 2006).

Acknowledgments

AKD thanks Vladimir Uversky for initiating these studies in his laboratory. This work was supported in part by Ministry of the Education, Culture, Sports, Science and Technology, Japan (21113007 to MO, HH and SF), National Institute of Health, USA (R01 GM071714-01A2) and National Science Foundation, USA (EF 0849803) (to AKD).

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.jsb.2012.10.017.

References

Altschul, S.F. et al., 1997. Gapped BLAST and PSI–BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25, 3389–3402.
Berman, H. et al., 2003. Announcing the worldwide Protein Data Bank. Nat. Struct. Biol. 10, 980.
Brunger, A.T., 1997. X-ray crystallography and NMR reveal complementary views of structure and dynamics. Nat. Struct. Biol. 4 (Suppl.), 862–865.
Cheng, Y. et al., 2006. Abundance of intrinsic disorder in protein associated with cardiovascular disease. Biochemistry 45, 10448–10460.
Chou, P.Y., Fasman, G.D., 1978. Prediction of the secondary structure of proteins from their amino acid sequence. Adv. Enzymol. Relat. Areas Mol. Biol. 47, 45–148.
Doreleijers, J.F. et al., 2012. NRG-CING: integrated validation reports of remediated experimental biomolecular NMR data and coordinates in wwPDB. Nucleic Acids Res. 40, D519–D524.
Dunker, A.K. et al., 2002. Intrinsic disorder and protein function. Biochemistry 41, 6573–6582.
Dunker, A.K. et al., 2001. Intrinsically disordered protein. J. Mol. Graph. Model. 19, 26–59.
Dyson, H.J., Wright, P.E., 2005. Intrinsically unstructured proteins and their functions. Nat. Rev. Mol. Cell Biol. 6, 197–208.
Feng, Y. et al., 2007. Solution structure and mapping of a very weak calcium-binding site of human translationally controlled tumor protein by NMR. Arch. Biochem. Biophys. 467, 48–57.
Fukuchi, S. et al., 2011. Binary classification of protein molecules into intrinsically disordered and ordered segments. BMC Struct. Biol. 11, 29.
Fukuchi, S. et al., 2012. IDEAL: intrinsically disordered proteins with extensive annotations and literature. Nucleic Acids Res. 40, D507–D511.
Fuxreiter, M. et al., 2004. Preformed structural elements feature in partner recognition by intrinsically unstructured proteins. J. Mol. Biol. 338, 1015–1026.
Galea, C.A. et al., 2008. Regulation of cell division by intrinsically unstructured proteins: intrinsic flexibility, modularity, and signaling conduits. Biochemistry 47, 7598–7609.
Garbuzynskiy, S.O. et al., 2005. Comparison of X-ray and NMR structures: is there a systematic difference in residue contacts between X-ray-and NMR-resolved protein structures? Proteins 60, 139–147.
Goroncy, A.K. et al., 2009. NMR solution structures of actin depolymerizing factor homology domains. Protein Sci. 18, 2384–2392.
Gould, C.M. et al., 2010. ELM: the status of the 2010 eukaryotic linear motif resource. Nucleic Acids Res. 38, D167–D180.
He, B. et al., 2009. Predicting intrinsic disorder in proteins: an overview. Cell Res. 19, 929–949.
Hellman, M. et al., 2004. Solution structure of coactosin reveals structural homology to ADF/cofilin family proteins. FEBS Lett. 576, 91–96.
Henrick, K., Thornton, J.M., 1998. PQS: a protein quaternary structure file server. Trends Biochem. Sci. 23, 358–361.
Hornbeck, P.V. et al., 2012. PhosphoSitePlus: a comprehensive resource for investigating the structure and function of experimentally determined post-translational modifications in man and mouse. Nucleic Acids Res. 40, D261–D270.
Iakoucheva, L.M. et al., 2002. Intrinsic disorder in cell-signaling and cancer-associated proteins. J. Mol. Biol. 323, 573–584.
Jones, D.T., Ward, J.J., 2003. Prediction of disordered regions in proteins from position specific score matrices. Proteins 53 (Suppl. 6), 573–578.
Larsson, G. et al., 2003. Detection of nano-second internal motion and determination of overall tumbling times independent of the time scale of internal motion in proteins from NMR relaxation data. J. Biomol. NMR 27, 291–312.
Lartigue, A. et al., 2002. X-ray structure and ligand binding study of a moth chemosensory protein. J. Biol. Chem. 277, 32094–32098.
Le Gall, T. et al., 2007. Intrinsic disorder in the Protein Data Bank. J. Biomol. Struct. Dyn. 24, 325–342.
Lee, M.R. et al., 2008. Inhibition of APP intracellular domain (AICD) transcriptional activity via covalent conjugation with Nedd8. Biochem. Biophys. Res. Commun. 366, 976–981.
Lobanov, M.Y. et al., 2010. Library of disordered patterns in 3D protein structures. PLoS Comput. Biol. 6, e1000958.
Markley, J.L. et al., 1998. Recommendations for the presentation of NMR structures of proteins and nucleic acids. IUPAC–IUBMB–IUPAB inter-union task group on the standardization of data bases of protein and nucleic acid structures determined by NMR spectroscopy. J. Biomol. NMR 12, 1–23.
Marx, U.C. et al., 2000. Solution structures of human parathyroid hormone fragments hPTH(1–34) and hPTH(1–39) and bovine parathyroid hormone fragment bPTH(1–37). Biochem. Biophys. Res. Commun. 267, 213–220.
Matthews, B.W., 1975. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. Biochim. Biophys. Acta 405, 442–451.
Minezaki, Y. et al., 2006. Human transcription factors contain a high fraction of intrinsically disordered regions essential for transcriptional regulation. J. Mol. Biol. 359, 1137–1149.
Mittag, T., Forman-Kay, J.D., 2007. Atomic-level characterization of disordered protein ensembles. Curr. Opin. Struct. Biol. 17, 3–14.
Mohan, A. et al., 2009. Influence of sequence changes and environment on intrinsically disordered proteins. PLoS Comput. Biol. 5, e1000497.
Mohan, A. et al., 2006. Analysis of molecular recognition features (MoRFs). J. Mol. Biol. 362, 1043–1059.
Mosbah, A. et al., 2003. Solution structure of a chemosensory protein from the moth *Mamestra brassicae*. Biochem. J. 369, 39–44.
Muchmore, S.W. et al., 1996. X-ray and NMR structure of human Bcl-xL, an inhibitor of programmed cell death. Nature 381, 335–341.
Pawley, N.H. et al., 2001. An improved method for distinguishing between anisotropic tumbling and chemical exchange in analysis of 15N relaxation parameters. J. Biomol. NMR 20, 149–165.
Rabut, G., Peter, M., 2008. Function and regulation of protein neddylation. 'Protein modifications: beyond the usual suspects' review series. EMBO Rep. 9, 969–976.
Serrano, P. et al., 2010. Comparison of NMR and crystal structures highlights conformational isomerism in protein active sites. Acta Crystallogr., Sect. F: Struct. Biol. Cryst. Commun. 66, 1393–1405.
Sickmeier, M. et al., 2007. DisProt: the database of disordered proteins. Nucleic Acids Res. 35, D786–D793.
Sikic, K. et al., 2010. Systematic comparison of crystal and NMR protein structures deposited in the protein data bank. Open Biochem. J. 4, 83–95.
Sprangers, R. et al., 2000. Refinement of the protein backbone angle PSI in NMR structure calculations. J. Biomol. NMR 16, 47–58.
Sugase, K. et al., 2007a. Mechanism of coupled folding and binding of an intrinsically disordered protein. Nature 447, 1021–1025.
Sugase, K. et al., 2007b. Tailoring relaxation dispersion experiments for fast-associating protein complexes. J. Am. Chem. Soc. 129, 13406–13407.
Tamiola, K. et al., 2012. Sequence-specific random coil chemical shifts of intrinsically disordered proteins. J. Am. Chem. Soc. 132, 18000–18003.
The UniProt Consortium, 2011. Ongoing and future developments at the Universal Protein Resource. Nucleic Acids Res. 39, D214–D219.
Tomaselli, S. et al., 2006. The alpha-to-beta conformational transition of Alzheimer's Abeta-(1–42) peptide in aqueous media is reversible: a step by step conformational analysis suggests the location of beta conformation seeding. ChemBioChem 7, 257–267.
Tompa, P., 2005. The interplay between structure and function in intrinsically unstructured proteins. FEBS Lett. 579, 3346–3354.
Tzakos, A.G. et al., 2006. NMR techniques for very large proteins and rnas in solution. Annu. Rev. Biophys. Biomol. Struct. 35, 319–342.
Ulrich, E.L. et al., 2008. BioMagResBank. Nucleic Acids Res. 36, D402–D408.
Uversky, V.N., Dunker, A.K., 2010. Understanding protein non-folding. Biochim. Biophys. Acta 1804, 1231–1264.
Uversky, V.N. et al., 2000. Why are "natively unfolded" proteins unstructured under physiologic conditions? Proteins 41, 415–427.
Ward, J.J. et al., 2004. Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. J. Mol. Biol. 337, 635–645.
Wright, P.E., Dyson, H.J., 1999. Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm. J. Mol. Biol. 293, 321–331.
Zhang, M. et al., 1995. Calcium-induced conformational transition revealed by the solution structure of apo calmodulin. Nat. Struct. Biol. 2, 758–767.
Zhang, Y. et al., 2007. Between order and disorder in protein structures: analysis of "dual personality" fragments in proteins. Structure 15, 1141–1147.