Published in final edited form as:

Biochim Biophys Acta. 2013 February ; 1834(2): 487–498. doi:10.1016/j.bbapap.2012.12.003.

Utilization of protein intrinsic disorder knowledge in structural proteomics

Christopher J. Oldfield^{a,*}, Bin Xue^b, Ya-Yue Van^c, Eldon L. Ulrich^d, John L. Markley^d, A. Keith Dunker^{a,**}, and Vladimir N. Uversky^{b,e,f,***}

^a Center for Computational Biology and Bioinformatics, Department of Biochemistry and Molecular Biology, Indiana University School of Medicine, Indianapolis, Indiana 46202, USA

^b Department of Molecular Medicine, College of Medicine, University of South Florida, Tampa, Florida 33612, USA

^c Molecular Kinetics, 6201 La Pas Trail, Indianapolis, Indiana 46268, USA

^d Center for Eukaryotic Structural Genomics, University of Wisconsin-Madison, 433 Babcock Drive, Madison, Wisconsin 53706-1549, USA

^e Byrd Alzheimer's Research Institute, College of Medicine, University of South Florida, Tampa, FL 33612, USA

^f Institute for Biological instrumentation, Russian Academy of Sciences, 142290 Pushchino, Moscow Region, Russia

Abstract

Intrinsically disordered proteins (IDPs) and proteins with long disordered regions are highly abundant in various proteomes. Despite their lack of well-defined ordered structure, these proteins and regions are frequently involved in crucial biological processes. Although in recent years these proteins have attracted the attention of many researchers, IDPs represent a significant challenge for structural characterization since these proteins can impact many of the processes in the structure determination pipeline. Here we investigate the effects of IDPs on the structure determination process and the utility of disorder prediction in selecting and improving proteins for structural characterization. Examination of the extent of intrinsic disorder in existing crystal structures found that relatively few protein crystal structures contain extensive regions of intrinsic disorder. Although intrinsic disorder is not the only cause of crystallization failures and many structured proteins cannot be crystallized, filtering out highly disordered proteins from structure-determination target lists is still likely to be cost effective. Therefore it is desirable to avoid highly disordered proteins from structure-determination target lists and we show that disorder prediction can be applied effectively to enrich structure determination pipelines with proteins more likely to yield crystal structures. For structural investigation of specific proteins, disorder prediction can be used to improve targets for structure determination. Finally, a framework for considering intrinsic disorder in the structure determination pipeline is proposed.

© 2012 Elsevier B.V. All rights reserved

* Correspondence to: C.J. Oldfield, Center for Computational Biology and Bioinformatics, Indiana University Schools of Medicine and Informatics, 410 W. 10th Street, Suite 5000, Indianapolis, IN 46202, USA. cjoldfie@uimail.iu.edu (C.J. Oldfield) . **

Correspondence to: A.K. Dunker, Center for Computational Biology and Bioinformatics, Indiana University Schools of Medicine and Informatics, 410 W. 10th Street, Suite 5000, Indianapolis, IN 46202, USA. Tel.: +1 317 278 9220; fax: +1 317 278 9217.

kedunker@iupui.edu (A.K. Dunker) . *** Correspondence to: V.N. Uversky, Department of Molecular Medicine, University of South Florida, 12901 Bruce B. Downs Blvd. MDC07, Tampa, Florida 33612, USA. Tel.: +1 317 278 6448; fax: +1 317 278 9217.

vuversky@health.usf.edu (V.N. Uversky) .

Appendix A. Supplementary data

Supplementary data to this article can be found online at <http://dx.doi.org/10.1016/j.bbapap.2012.12.003>.

Keywords

Proteomics; Structural genomics; Structural proteomics; Intrinsically disordered protein

1. Introduction

Structure determination, historically, has been attempted on a protein-by-protein basis, typically after an accumulation of years or decades of study on each particular protein. Information regarding solubility, stability, pH range, and temperature sensitivity was therefore generally well known. Consequently, due to lack of suitable samples, structure determination was not usually attempted on proteins that were ill behaved in solution. The Protein Structure Initiative (PSI) or Structural Genomics Initiative (SGI) [1-4] has turned the *status quo* on its head by attempting to determine structures without prior knowledge of a protein's behavior and to do this rapidly on a large scale. Despite the obvious success of the PSI Centers in decreasing the overall cost of determining novel structures [5], ill-behaved proteins continue to represent major challenges that hamper the efficiency of structure determination. One source of ill-behaved proteins is intrinsic protein disorder.

Prediction techniques and data mining have shown that intrinsically disordered proteins (IDPs) and proteins with regions of intrinsic disorder are likely to be quite common [6-12]. Disordered proteins or regions are defined herein as entire proteins or regions of proteins that lack a fixed tertiary structure. A given region of intrinsic disorder might be ordered or disordered, depending on the physiological or experimental conditions. For instance, a protein may undergo a disorder-to-order transition upon binding to a cofactor, DNA, or protein partner. On the other hand, proteins that lose rigid structure in the presence of denaturants are not considered to be disordered proteins. Similarly, proteins that fold into specific 3-D structure only under condition of extreme molecular crowding are not considered ordered. IDPs and intrinsically disordered regions are noticeably different from structured globular proteins and their domains at several levels, including amino acid composition, sequence complexity, hydrophobicity, charge, and flexibility. Many of these differences were used in the development of several disorder predictors (reviewed in Refs. [13-18]).

Disordered protein can impact many of the processes in the structure determination pipeline, including expression and stability [19,20], solubility [21,22], and crystallization [22-27]. Therefore, it is advantageous to filter highly disordered proteins from the target list. Also, many structured proteins contain isolated regions of intrinsic disorder, which can inhibit crystallization. Fragments remaining after removal of disordered regions may crystallize when the whole protein did not [22], and so a method for identification of disordered regions to allow for intelligent target improvement would be of great utility.

Here we examine the impact of intrinsic disorder on the various stages of the structure determination pipeline and outline several applications of Predictors of Natural Disordered Regions (PONDR®s) that can improve the efficiency of structure determination efforts. To evaluate the tolerance of intrinsic disorder in protein crystals, missing density in the Protein Data Bank (PDB) was evaluated in light of the various sources of missing density in addition to intrinsic disorder. This analysis indicates that extensive disorder in PDB is relatively rare and highlights the benefit of filtering crystallization targets for intrinsic disorder. Several disorder-based target prioritization criteria are evaluated and retrospectively applied to protein targets in the TargetTrack database, the progress tracking module of the PSI Structural Genomics Knowledgebase. This evaluation suggests that disorder prediction provides an effective means for prioritizing targets for structure

determination pipelines. Finally, the use of disorder prediction for tailoring proteins for structure determination is examined.

2. Materials and methods

2.1. Protein datasets

For analysis of missing density, protein structures without nucleic acid from the July 2012 version of PDB [28] were used in conjunction with the S2C database (<http://dunbrack.fccc.edu/Guoli/s2c/index.php>). The latter provides alignments between residues with defined density and the reported sequence (SEQRES), greatly simplifying identification of apparent missing density. The number of residues and proportion of apparent missing density were calculated with respect to the contents of the asymmetric unit. Several structures were excluded from the list of cases of extreme missing density due to idiosyncratic issues with these structures: 1EFM because a large portion of defined density was not fit by the author [29], 1YD7 because no supporting reference was available and neither solvent content nor Mathew's coefficient was reported, and 3JQH and 4HB1 because these structures contain chains with exact repeats and the asymmetric unit contains less than one copy of the chain in the asymmetric unit.

Three idealized sets of proteins were assembled for these studies. The "ordered proteins" set consists of the entire sequences of all the X-ray crystallography-determined structures from PDB Select 25 proteins [30]. This set contains 929 proteins with 230,205 residues. The "disordered proteins" set consists of proteins compiled from literature searches where the known disordered regions comprise 80% of the entire sequences. This set contains 48 proteins with 9,219 residues. The "X-ray proteins with disorder" set consists of crystallized proteins with disordered regions 30 residues. This set contains 53 proteins with 17,191 residues.

We also analyzed all the proteins currently available at the PSI Structural Biology Knowledgebase (<http://sbkb.org>), namely at its target progress tracking module (TargetTrack: <http://sbkb.org/tt/>), which provides status and tracking information on the production protein targets and their structure solution [31]. Targets' sequence and status were taken from the XML distribution of the database, as of July 2012. Status tags for each trial were normalized by adding tags from previous pipeline steps that were missing. For targets with multiple trials, the trial showing the furthest pipeline progression was selected as representative of the target. From 302,311 selected targets deposited by worldwide Contributing Centers in to TargetTrack, 201,825 were cloned, of which 120,802 were expressed. Of the expressed proteins, 51,303 were purified, of which 8,168 were crystallized, of which 5,125 produced diffraction-quality crystals. So far, according to the July 2012 TargetTrack, 4,203 crystal and 1,600 NMR structures have been solved.

2.2. Intrinsic disorder prediction

PONDR® algorithms predict per-residue, two-state order/disorder from amino acid sequence [32-34]. Although several PONDR®s have been developed to date [9,32-40], only one of these predictors, VL-XT, is used here. This choice of this predictor was made because PONDR® VL-XT is one of the best characterized members of the set of PONDR® predictors.

PONDR® predictions are real numbers between 0 and 1, where values of 0.5 or more are predictions of disorder. Raw PONDR® prediction output is commonly plotted, as distribution of disorder scores over the protein sequence. An alternative representation is to map predictions to a bar, where black represents a region predicted to be disordered and white represents a region predicted to be ordered. Predictions can also be quantified per-

protein, using measures such as percent of residues predicted to be disordered or the longest predicted disordered region. These measures are particularly useful because of the dependency of false positive rates on the length of the disorder prediction.

PONDR® VL-XT was constructed from a neural network trained on disordered proteins collected from the literature. The VL-XT predictor integrates three feed-forward neural network predictors: the various long predictor version 1.0 (VL1) [36], the N-terminus predictor (XN), and the C-terminus predictor (XC) [33], and achieves good accuracy both on field-out training set sequences and on out-of-sample sequences [9]. Prediction of VL-XT on O_PDB_S25 [9], which is a set of defined density regions from sequence unique structures from PDB, gave an accuracy of 80%. Prediction on the disordered data gave an accuracy of 60%, suggesting that VL-XT does not generalize well for prediction of disorder [9].

PONDR® is available on the web at <http://www.pondr.com> (use is free, but registration is required). The web interface does not support batch predictions. Arrangements for large numbers of predictions can be made by contacting the corresponding authors. It is important to note that although there are many predictors of protein intrinsic disorder [13-18], we anticipate that similar results would be found with other predictors.

2.3. Model fitting

To evaluate combinations of disorder prediction based criteria, logistic regression models were fit using proteins that reached a given status in the structure determination pipeline as a positive set and remaining proteins as a negative set. The purpose of fitting these models was not to create binary predictors of status progression, but rather to determine an appropriate linear combination of parameters. The statistics package R (<http://www.r-project.org/>) was used to fit models, using the standard generalized linear model function to fit models and the package ROCR [41] to calculate the area under the curve (AUC) measure of predictor performance. The AUC was selected to compare models because it emphasizes the separation of the positive and negative dataset, where 1.0 is perfect separation and 0.5 is completely interspersed, and is independent of any kind of prediction threshold.

3. Results and discussion

3.1. Missing density and intrinsic disorder in the PDB

To evaluate the tolerance of crystal structures for intrinsic disorder, we examined the extent of missing density in existing crystal structures. Although missing density regions have often been equated to intrinsic disorder, apparent missing density in crystal structures may arise from several sources, including disordered regions; mobile domains, and proteolysis. Disordered regions may be present in the crystal but fail to diffract X-rays due to heterogeneous or dynamic structure. For multi-domain proteins, mobile domains may be present; a domain may have a heterogeneous or dynamic orientation and position in the crystal, resulting in missing density in the solved crystal structure. Although not intrinsically disordered themselves, the mobility of domains may be facilitated by intrinsically disordered linkers. Finally, rational or fortuitous proteolysis may occur prior to or during crystallization, but these altered sequences are often not reflected in the PDB record. While several authors have noted that removal of intrinsically disordered regions can facilitate crystallization and diffraction, e.g. Ref. [22], the use of proteolysis is not consistently reported and the integrity of crystallized chains not routinely verified.

The most extreme cases of apparent missing density in terms of longest region of missing density and largest fraction of residues in missing density regions were collected from PDB (Table 1). All of these structures have over half of their residues in apparent missing density regions or at least one apparent missing density region longer than 200 residues. Much of

the missing density found was due to the anticipated sources, (likely) proteolysis, (likely) mobile domains and intrinsic disorder, but much apparent missing density was due to misannotation. Misannotations are less frequent than other sources of missing density and are limited to cases where only a portion of a sequence was included in experiments, but a longer sequence was reported in the PDB record, e.g. PDB ID: 4ACO. For the purposes of this study, apparent missing density due to misannotations is functionally equivalent to proteolysis. Discounting misannotation and proteolysis, and PDB record limitations, missing density regions attributable to residues physically present in protein crystals are limited to about 50% of residues when mobile domains (around 40% of residues for confirmed cases of mobile domains) are present and about 30% of residues for intrinsically disordered regions.

In addition to extreme cases of missing density in PDB, missing density across crystal structures in PDB was evaluated (Fig. 1). To account for the presence of mobile domains, we applied disorder prediction to the sequences of all PDB chains and filtered missing density regions by the fraction of residues predicted to be disordered. However, we were unable to account for proteolysis or misannotation in this analysis, so these results should be considered an upper bound; there is less missing density due to residues physically present in the protein crystal than indicated by these results. The cumulative fraction of missing density regions by length (Fig. 1A) shows that missing density regions are often short, with between 85% and 95% of regions being 20 residues long or shorter depending on the predicted disorder cutoff. Also, missing density regions with a moderate fraction of predicted disorder (0.3) tend to be longer than all missing density regions and regions with a larger proportion of predicted disorder, which suggests that many of these intermediately disordered regions may include mobile domains. The cumulative fraction of structures by length of its longest missing density region (Fig. 1B) shows that between 23% and 47% of structures have no missing density residues and that the vast majority of structures, between 85% and 95%, contain no missing density regions longer than 20 residues.

The overall picture of missing density in the PDB provided by this analysis is that missing density is infrequent in PDB structures, and when present tends to be limited to relatively short regions. When long regions of disorder are present in the crystal (Table 1, structures 1M1J, 3KIC and 3NG9) the total proportion of disordered residues in the structure is below 30%. Disorder prediction suggests that much of the missing density in PDB does arise from intrinsically disordered regions, while some arises from mobile domains and an unknown amount from proteolysis and misannotation. The limited amount of intrinsic disorder present as missing density regions agrees with the idea that intrinsically disordered regions, particularly long disordered regions, inhibits successful determination of crystal structures, and suggests that avoiding or tailoring disordered proteins may aid in the determination of crystal structures.

3.2. Target prioritization with disorder prediction

Per-protein assessments of disorder by PONDR® can contribute to a target prioritization scheme in a straightforward manner. For example, a cut-off value for a given quantity establishes a filter, while weights determined from the given measure provide a means for prioritization. Using different scoring schemes, specific cut-off values are suggested below, but it is suggested that alternative values be considered for each particular application in order to balance the rates of false prediction of disorder with tolerable disorder content. Likewise, the relative weights calculated from the different disorder schemes could be used to prioritize the selected targets with the proteins having less predicted disorder receiving the higher ranks.

The genome-wide measures of predicted disorder will be effective in prioritization, provided proper attention is paid to the balance between false prediction of disorder and false prediction of order. As illustrated in Table 2, the rate of false prediction of disorder drops drastically with increasing number of contiguous predictions of disorder. However, there is a corresponding increase in false prediction of order. There is no clear indication of the extent of this increase, due to the limited number of proteins in the current disorder protein dataset and uncertainty in the boundaries of disordered regions. Prediction sensitivity will be most directly affected by limiting the minimum size of disordered regions that prediction will detect. Because many procedures in molecular biology, including protein crystallization, can be relatively insensitive to small regions of disorder, ignoring contiguous disorder predictions of modest length (e.g., less than 40 residues) should not adversely affect the practical use of PONDR® predictions. The thresholds for any of the weighting schemes presented should be selected, in reference to Table 2, so that the overall rate of false prediction of disorder is in balance with the desired preference for the sensitivity of the predictions. Finally, it should be noted that the rate of false prediction of disorder may be somewhat overestimated in Table 2. The 0_PDB_S25 contains chains from complex structures, which may contain disorder-to-order transition regions involved in binding [6] that are correctly predicted to be disordered.

3.2.1. Target prioritization using percent predicted protein disorder—Shown in Fig. 2A are the cumulative histograms for the ordered, disordered, and X-ray with disorder datasets as a function of the percent of total sequence predicted to be disordered. This measure gives good separation between the sets of structured proteins and disordered proteins and could serve to indicate proteins that are mostly disordered. The weakness of this approach is that, while proteins that are highly disordered can be filtered, mostly ordered sequences containing enough disorder to hinder high throughput protocols may be overlooked by this measure. However, if longer proteins receive a lower priority, a simple percent of residues predicted to be disordered cutoff may be useful without additional criteria.

The percent of proteins with proportions of predicted disorder less than that indicated on the X-axis rises much more sharply for the ordered proteins and the X-ray proteins with disorder sets than the disordered protein set (see Fig. 2A). Although such a relationship is expected for an order-disorder predictor of reasonable accuracy, the position of the X-ray disorder set is particularly important. A large fraction of proteins in many genomes likely contain relatively long regions of disorder [7,8,10,12], and such proteins should be removed from the structure determination target list. However, the presence of relatively short regions of disorder does not imply that a protein would not be amenable to structure determination. Since a high proportion of X-ray disorder proteins have relatively low predicted disorder content, a prioritization scheme using the percent of residues predicted to be disordered would be useful.

Two example cut-off values are indicated in the cumulative histograms of Fig. 2A. The dotted lines in the figure illustrate an aggressive filtering and weighting scheme using 30% or less of residues predicted to be disordered as the cutoff. Based on these example datasets, this would eliminate 90% of the disordered proteins while retaining 75% of the ordered proteins and 55% of the crystallizable proteins having disordered regions. A more conservative cutoff is illustrated with the dashed lines in Fig. 2A. Using this threshold of 60% or less of each sequence predicted to be disordered, nearly all of the ordered proteins and the X-ray disordered proteins (98% and 96%, respectively) can be retained while eliminating over half (56%) of the disordered proteins.

3.2.2. Target prioritization using total predicted protein disorder—For larger proteins, a small percentage of predicted disorder could correspond to a large amount of total disorder, which could inhibit the structure determination process. Two alternative measures of total predicted disorder are suggested as possibilities to overcome this problem (see Fig. 2B and C). Proteins removed from a target list by these two measures could be separated into their ordered regions and their disordered regions for further study.

The two suggested measures of total predicted disorder are analyzed in Fig. 2B and C. The first measure, Fig. 2B, considers the longest disorder prediction in each protein, where the length is given by the x-axis value. The second measure, Fig. 2C, presents the number of residues in predicted regions of disorder of 20 residues or longer in each protein, where the number of residues is given by the x-axis value.

The longest predicted disordered region could be used for prioritization by eliminating proteins with predicted disordered regions of 40 residues or longer. Such an aggressive scheme would retain 91% of ordered proteins and 68% of X-ray disorder proteins as targets, and eliminate 58% of disordered proteins from consideration, as indicated by the dotted line in Fig. 2B. A longest predicted disordered region cutoff of 70 residues, represented by the dashed line in the same figure, represents a more conservative filtering scheme. This would retain 99% of ordered proteins and 89% of X-ray disorder proteins but eliminate only 29% of completely disordered proteins.

Alternatively, the total number of residues in predicted disordered regions of 20 or longer could be used as a prioritization attribute, with a cutoff of 40 residues as indicated by the dashed line in Fig. 2C. Approximately 62% of disordered proteins could be filtered while retaining 78% of ordered proteins and 59% of X-ray disorder proteins. With a conservative threshold of 50 residues, 94% of ordered proteins and 76% of X-ray disorder proteins are retained while 44% of the disordered proteins would be eliminated.

3.3. Retrospective evaluation of predicted disorder based filtering

The effectiveness of the proposed methods of disorder-based filtering for crystallization targets was examined through a retrospective evaluation. For this, disorder predictions were made and evaluated for all proteins in the TargetTrack database. Since the use of disorder-based filtering in selected proteins would confound retrospective evaluation, we examined the TargetTrack database for biases in the content of predicted disorder. Based on the idea that targets were not filtered for disorder in the initial years of the PSI, targets were grouped by the year they were first selected and the values of the filtering criteria were compared. As can be seen by the plot of mean percent predicted disorder by year (Fig. 3), targets selected in early years of the project have significantly more disorder than those selected in the latter years of the project. Note that the high proportion of predicted disordered in targets selected in 2004 and prior relative to those selected in 2007 and later is not reflected in analogous plots for total disorder (Fig. S1 and S2). The latter plots show no pattern in the bias of disorder content based on the selection year of targets. These results suggest that disorder prediction has been used for target filtering by contributing centers using a criterion similar to percent predicted disorder. Based on this result, targets selected in 2004 or before were taken as a set of unfiltered structure targets, and targets selected in 2007 and before were taken as a set of filtered targets.

Disorder prediction based filtering was evaluated by CDF plots for unfiltered targets using the proportion predicted disorder (Fig. 4A), number of residues in disordered regions of 20 residues or longer (Fig. 4B), and longest disordered region (Fig. S3) criteria. These plots have one curve for each target status in the crystal structure pipeline, which progresses in the following order: selection, cloning, expression, purification, crystallization, diffraction, and

structure determination. Each criterion shows a progressive loss of disordered proteins in subsequent steps of the structure determination pipeline. Also, the curves for targets with determined structures (black line in Figs. 4A, B, and S3) closely follow the curves for ordered proteins (empty line in Fig. 2A, B, and C), which indicates that ordered proteins are an appropriate model for successful structure targets. Comparison of the curves for selected targets with curves for targets with determined structures suggests that disorder based filtering provides an effective method for increasing pipeline efficiency; depending on criteria and threshold selection, 5% to 20% of selected proteins could be filtered at the suggested thresholds while losing relatively few successful structure targets.

The impact of disorder based filtering was also investigated by examining filtering criteria for the set of filtered targets, i.e. those targets selected in 2007 and later. As expected, the plots of proportion of predicted disorder (Fig. 4C) show little to distinguish selected targets from successful structure targets. Therefore, little efficiency could be gained from further filters based on this criterion. However, both total disorder criteria, number of residues in LDRs (Fig. 4D) and longest disordered region (Fig. S4) show a large difference in predicted disorder between selected and successful structure targets, although this difference is somewhat reduced compared to unfiltered targets (Figs. 4 and S3, respectively). This suggests that not only is a low proportion of intrinsic disorder important for successful structure determination, but also that there is a limit to the absolute number of disordered residues a well-diffracting crystal can tolerate. Therefore, further gains in pipeline efficiency could possibly be gained by the use of multiple criteria.

Combinations of disorder criteria were examined by fitting logistic regression models and evaluating these models with the AUC performance measure. Single parameter models were also evaluated for comparison. These models were fit on proteins selected in 2004 and prior from the TargetTrack database, using proteins yielding crystal structures as one class and all other proteins as the other class. The goal in fitting these models was not to develop binary predictors of whether or not a protein will yield a crystal structure, but rather to determine an appropriate linear combination of parameters. AUC is used as a measure of separation between the two classes, which is appropriate since AUC is independent of threshold and the logistic function transform.

Comparison of AUCs for combinations of parameters (Table 3) shows that addition of parameters improves separation performance. Combination of all three disorder prediction parameters provides the best performance. However, all three parameters is only marginally better than the next best combination, fraction of disordered residues (FDR) and number of residues in long disordered regions (NDR), and the latter is examined further. The cumulative histogram of the linear combination of FDR and NDR is shown in Fig. 5. Note that the linear combination is $0.0092 \text{ NDR} - 0.72 \text{ FDR}$, where the signs indicate that NDR has a negative impact on crystallization but that FDR has a positive impact on crystallization. The apparent positive influence of FDR on pipeline progress, which contradicts the current working hypothesis, was investigated further by fitting FDR-NDR logistic regression models for each combination of subsequent target status tags (Table S1). The positive relationship between FDR and structure pipeline progression persists except for progression of purified targets, for which both FDR and NDR are detrimental.

The relationship between predicted disorder and structure pipeline progression suggested by these results is 3-fold: (1) NDR generally inhibits pipeline progression, while (2) FDR (in conjunction with NDR) benefits pipeline progression from expression through purification but (3) FDR (in conjunction with NDR) inhibits pipeline progression from purification through to crystallization. Relationship (2) is somewhat contrary to the current working hypothesis, but can be reconciled in the following way. Successful protein purification is

dependent on protein solubility, so the current observation is consistent with recent results that show that intrinsic disorder can greatly increase protein solubility [42]. Relationships (1) and (3) agree with expectation since disorder has been observed to complicate each step in the structure determination process, from expression to solubility to crystallization.

Expression levels can suffer from the reduced stability of intrinsically disordered proteins, due to their inherent proteolytic sensitivity; intrinsically disordered proteins are digested much faster than ordered proteins [43-47]. Solubility of some proteins with disordered regions can be very low under conditions promoting the native state. For example, the aggregation of wild type Bcl-2 was remedied by replacing its disordered region of Bcl-2 with a portion of the disordered region of Bcl-xL, reducing aggregation and allowing structure determination of the chimera [21]. Crystallization is unlikely for completely disordered proteins or proteins with too much intrinsic disorder because of their flexible and very dynamic nature. Myelin basic protein (MBP) exemplifies these troublemakers [48]. One exhaustive series of attempts to crystallize MBP for X-ray diffraction has been reported, where the authors tried 4,600 different crystallization conditions but were unable to induce crystallization of MBP [25]. Based on these observations the myelin basic protein has been suggested to belong to the category on “uncrystallizable” proteins. It can be safely assumed that many other unsuccessful crystallization attempts for numerous other proteins have not been reported, since negative results are generally assumed to be unsuitable for publication. In the case of MBP, several additional studies suggest that this protein lacks fixed 3D structure, existing instead as an intrinsically disordered ensemble, which in turn have been suggested to provide the basis for its multifunctionality [49].

3.4. Target improvement with disorder prediction

Many protein structures in PDB are fragments of full-length gene products. These structures often represent independently folding domains and are invaluable for understanding the biological functions of these proteins. Ordered regions need to be identified and isolated, preferably without the use of extensive (and costly) experimentation. This section illustrates some ways in which expert analysis of PONDR® predictions can provide a starting point for the isolation of ordered regions and disordered regions from their larger host proteins.

3.4.1. Disordered termini detection—Many proteins contain long regions of disorder at either the amino or carboxyl terminus or at both termini. Removing these terminal disordered regions can often improve protein handling and crystallizability. Crystallization trials for three proteins, NEIL1 [22], RluC and β -recombinase, exemplify such a target improvement scheme. The original crystallization attempts for these proteins began with the full-length sequences.

NEIL1, a human homolog of *Escherichia coli* DNA glycosylase endonuclease VIII, represents an illustration of the non-trivial correlation between disorder, solubility, and crystallizability. Crystallization trials for full-length C-terminally His-tagged NEIL1 failed to yield any crystals. This inability to grow crystals was corroborated by the fact that the protein was polydisperse based on the dynamic light-scattering analysis, and this polydisperse feature persisted regardless of the temperature or buffer conditions used [22]. To overcome this obstacle, PONDR® VL-XT was used to detect disordered region(s) in NEIL1 that may hinder crystallization. This analysis predicted that the protein has a disordered C-terminal region [22]. Based on this observation, a construct missing the entire disordered C-terminal region (last 106 amino acids) in NEIL1 was designed. Little or no soluble protein expression was observed with this construct. A series of shorter C-terminal deletion constructs were cloned and checked for expression. These studies showed that deletions of > 100 amino acids did not yield any protein expression. Therefore, a NEIL1

construct missing the C-terminal 100 amino acids (NEIL1CΔ100) was tested in crystallization studies. This active construct was successfully overexpressed in *E. coli*, purified and crystallized [22]. Thus, this example clearly illustrates the potential utility of disorder prediction to increase chances for successful protein crystallization.

Crystals were obtained for RluC after a 6-week crystallization period and the protein in the crystal was found to lack 88 residues from the amino terminus. Further analysis showed that trace protease activity was likely responsible for the hydrolysis of the backbone. Further crystallization trials using truncated RluC succeeded in 2–3 days [50]. PONDR® predicts the removed region to be highly disordered, with a disorder–order boundary within 2 residues of the observed cleavage site. Fig. 6 shows this truncation from two perspectives: Fig. 6A is a plot of the PONDR® prediction with the removed region indicated and Fig. 6B gives the order–disorder prediction in the form of a bar with the crystallized region as indicated. This bar-type representation is indicated here for later use.

Crystallization of full-length β-recombinase required 6 months to produce small crystals. Analysis of the protein in the crystal revealed that 82 residues at the carboxyl terminal had been cleaved [51]. The removed region is predicted to be highly disordered by PONDR®, with an order–disorder boundary within 10 residues of the cleavage site (data not shown).

The removed regions of RluC and β-recombinase have not been convincingly characterized as completely disordered. However, the extreme protease sensitivity of disordered regions suggests that these two proteins contain disorder, at least in the regions surrounding their cleavage sites. Both of the cleaved loci are predicted to be disordered by PONDR® VL-XT. Although neither of the termini of either protein is predicted to be completely disordered, removal of the bulk of the predicted disorder led to crystallization of both proteins.

In practice, the fraction of residues predicted to be disordered at each terminus should be considered, not just long predictions of disordered at the termini. If the terminal disorder represents a large portion of the predicted disorder in a protein, it is likely that removing such a region will improve a protein's crystallizability.

3.4.2. Detection of ordered fragments—Disorder prediction can be useful to identify the specific regions of proteins that may form rigid structures. Fig. 7 provides examples showing the relationship between PONDR® predictions (upper bar) and the locations of structured fragments (lower bar) for a representative set of proteins.

The predictions are obviously not perfect. Short regions of predicted disorder are contained within ordered regions, and the fragment boundaries are often over- or under-predicted. Such inaccuracies can be partially attributed to disordered regions that are misclassified as ordered regions, as exemplified the c-fos and c-jun entries.

Many ordered protein regions in PDB are involved in protein–protein or protein–nucleic, acid complexes, yet these same regions are disordered in the absence of their partners. By our definition, such regions should be classified as disordered, but such regions are typically misclassified as ordered in current data sets. This adds noise to the data, which contributes to the observed error rates. For example, DNA binding regions of proteins are often predicted to be disordered [7]. For some of these proteins experimental evidence has shown that such regions are disordered in the absence of DNA. An example of this type is the c-fos/c-max leucine zipper DNA binding region. Leucine zippers have been shown to be disordered in isolation from their zipper partner [52]; therefore these regions of c-fos and c-jun are classified as disordered. The 1FOS crystal structure contains the leucine zipper dimerization regions of c-fos and c-jun, which are complexed to DNA. The portions of these helices that

make contact with the DNA are isolated from any other protein contacts, which suggest that these regions would be highly unstable in the absence of DNA, and indeed, these regions were predicted to be disordered. Similar behavior was also found for the c-Myc/Max leucine zipper DNA binding region. Furthermore, although c-Myc and Max were characterized by the well-defined, ordered structure in their tertiary complex with DNA (1NKP), the detailed structural characterization of unbound Max and c-Myc revealed that both proteins were highly disordered in their free forms and underwent mutual coupled binding and folding when their zipper domains interacted to form a helical coiled coil (Fig. 8) [53,54]. These examples show that comprehensive study of disorder predictions for regions involved in protein–ligand interactions may provide useful information for future attempts at identifying partners of intrinsically disordered protein regions.

These results point to the importance of specific investigations of order–disorder boundaries to determine whether their prediction can be improved. However, even without further refinement, disorder prediction can at least provide a useful starting point for structure studies by indicating possible locations of the ordered regions of proteins.

3.4.3. Inclusion of binding partners—While this paper has focused on the inhibition of protein structure determination by excessive amounts of intrinsic disorder, there are counter examples that do not fit this general trend, including structured proteins that do not form suitable crystals and disorder-containing proteins that do form suitable crystals. Both of these features are discussed below in terms of a specific example.

When a protein fails to crystallize or gives poor quality crystals, a common approach is to attempt to crystallize a homologue, often leading to a successful end result. An interesting example is provided by attempts to determine the structure of protein phosphatase 1 (PP1) complexed with inhibitor 2 (12). When rabbit muscle PP1 was complexed with the rabbit 12 protein, crystals that diffracted to ~ 4 Å were obtained (Tom Hurley, personal communication). However, when experiments were carried out using rat PP1 complexed with mouse 12, crystals were also obtained, but these diffracted to 2.5 Å, which was sufficient for 3D structure determination [55].

The PP1 :12 complex also illustrates that proteins with substantial amounts of disorder can sometimes form crystals of sufficient quality for structure determination. When not bound to PPI, the 12 protein is unstructured [55,56]. Upon complex formation with PP1, three separate segments of 12 become structured, giving a total of 59 structured residues but with 154 residues that remain unobserved in the complex. An NMR study of PP1 complexed with a large 12 fragment, from residue 9 to 169, shows that 100 residues corresponding to two unobserved segments in the crystal indeed remain as a random coil in the complex. Furthermore, large holes of sufficient size to accommodate models constructed for the IDP ensemble are observed for the PP1:12 crystal lattice [57]. This crystal structure is similar to those presented in Table 1, where a large disordered region can be accommodated by the crystal lattice if a sufficient number of ordered residues are present to support the lattice.

3.5. Integration of disorder prediction with high throughput structure determination

One use of PONDR® in structural genomics would be to divide the target list into two separate pipelines: a structure pipeline, for whole proteins or protein domains for which standard structure determination would be appropriate (i.e., for highly ordered protein); and a disorder pipeline, for proteins or protein domains for which other methods of study would be appropriate (i.e., for highly disordered proteins). Although a structure pipeline represents the current Structure Genomics methodologies, a disorder pipeline has not been implemented as of yet.

The underlying rationale for the PSI is that protein function can be inferred through knowledge of protein structure [1,2]. Disordered proteins, however, are not encompassed by this view of protein function, including traditional ideas regarding protein flexibility, e.g., allostery and induced fit. While much information can be gathered from disordered proteins through biophysical and biochemical methods, new models are required to translate this information into functional mechanisms. The first step in the development of these new models is the accumulation of biophysical parameters from a broad range of disordered proteins. Through a centralized dataset of such information (e.g., DisProt database [58] or similar resources), generalities may emerge that will enable functional inferences, as is now possible with ordered proteins. The lack of this accumulated information on disordered protein is a major block to understanding the mechanisms of disordered protein function.

Disordered proteins can be analyzed by traditional biophysical methods including circular dichroism, infrared spectroscopy, fluorescence, various hydrodynamic studies, small-angle X-ray or neutron scattering, etc. Studies using these methods provide a picture of the average conformational range of proteins (i.e., such methods when used in combination differentiate extended and collapsed disorder) [47,59-62]. Mapping protein disorder by the means of limited proteolysis resolved through both SDS-PAGE and mass spectroscopy, which were successfully used to map disordered regions in XPA [24] and clusterin [23], would also be a generally applicable methodology. These mapping studies localize disordered regions in terms of protein primary structure and can indicate the presence of structured domains, as in XPA [24], or indicate the presence of regions of secondary structure, as in clusterin [23]. A more complete picture of the conformational dynamics of disordered proteins could be obtained through the use of small molecule probes and cross-linkers that would indicate the time averaged spatial relationships of individual residues, although these techniques have yet to be explored in this context.

A survey of disorder-function relationships for more than 90 proteins with experimentally determined regions of disorder of length ≥ 30 residues ascribed more than 25 separate functions to the regions of disorder, with these functions falling into 4 broad categories: 1) molecular recognition, 2) molecular assembly, 3) protein modification, and 4) entropic chains [6]. The molecular recognition category is especially interesting. Molecular recognition by ordered proteins typically involves catalysis that can be described by the lock-and-key paradigm, whereas molecular recognition by disordered proteins typically involves signaling with a disorder-to-order transition playing a key role. Given that disordered regions have been implicated as the primary interaction regions in signaling events [7,63-68], these regions of proteins are prime candidates for use in proteomic binding assays. Using protein chip technology [69], display of intrinsically disordered protein segments for screening against a host of gene products could prove to be advantageous in elucidating the web that underlies cell signaling. Thus, armed with biophysical and functional evidence, large advances in the understanding of disordered protein function could be made.

Disorder prediction can be used to differentiate proteins that are likely to crystallize from proteins that are unlikely to crystallize or that are almost certain to not crystallize. Proteins predicted to contain both ordered and disordered regions can be studied in both pipelines, using disorder prediction as a starting point for any modification necessary to facilitate analyses (e.g., removal of disorder from structured domains and isolation of disordered regions). Disorder prediction would also provide the basis for helping to identify structured and unstructured domains within multidomain proteins; such identified domains could then be recovered by genetic engineering and sent to their respective pipelines.

The structure pipeline and disorder pipeline would collect ordered structure information and disorder structural information in complementary databases that would allow for concurrent study of some proteins in both contexts for the greatest possible efficiency and expediency. Regardless of whether an approach such as this is used, disorder prediction could aid selection and improvement of protein targets for the structure-determination pipeline through the filtering of disordered protein targets and isolation of ordered domains.

4. Conclusions

In this study, the impact of intrinsic disorder on the structure determination pipeline was examined. Intrinsic disorder is tolerated to some extent in crystal structures, although extensive regions of intrinsic disorder are infrequent and are likely to require a sufficient relative proportion of ordered residues to support the crystal lattice. In fact, as shown in Table 1 and the associated references, large regions of apparently missing electron density in X-ray-determined protein structures in the PDB can arise from true intrinsic disorder, from mobile domains, from misannotation, or from unreported proteolysis, with intrinsic disorder accounting for about 30% of the mass of the protein in crystals containing the largest amounts of disorder. Given the high solvent content (~50%) of typical protein crystals (Tom Hurley, personal communication), the capacity of protein crystals to accommodate such large amounts of disorder is not surprising. This accommodation of disorder and the inaccuracy of disorder prediction (Table 2) combine to limit the ease of use of disorder prediction for filtering structural proteomic target lists. Despite these limitations, the use of intrinsic disorder prediction for filtering targets is a viable method to improve the proportion of ordered targets in the structure determination pipeline. Retrospective application of these methods additionally demonstrates that the effect of intrinsic disorder on structure determination is not only detrimental; increased success in purification is correlated with a higher proportion of predicted disordered residues, which agrees with recent observations [42]. For structure determination of individual proteins, prediction of intrinsic disorder can be an effective tool for tailoring proteins for structure determination. Though not developed to the extent to which automated analysis would be possible, consideration of prediction results on a protein-by-protein basis is certainly feasible. For the time being, these methods will likely prove useful in a supplemental capacity. That is, when difficult proteins are encountered, their PONDR® predictions can be consulted to determine if they suggest an ordered region that might crystallize or a disordered terminus that could be easily removed. In this way, PONDR® predictions could be used to salvage regions of proteins for the structure pipeline and provide input for disorder analysis. Finally, while this work focuses on the use of disorder prediction to facilitate structure determination, complementary approaches can be used to study those proteins and regions that fall outside the purview of direct structure determination methods.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work was supported in part by the National Institute of Health (NIH grant 1 P50 GM64598-01 to E.L.U. and J.L.M.), the National Science Foundation (NSF grant EF 0849803 to A.K.D and V.N.U.), and the Program of the Russian Academy of Sciences for the “Molecular and Cellular Biology” (to V.N.U.).

References

1. Burley SK. An overview of structural genomics. *Nat. Struct. Biol.* 2000; 7(Suppl.):932–934. [PubMed: 11103991]

2. Chance MR, Bresnick AR, Burley SK, Jiang JS, Lima CD, Sali A, Almo SC, Bonanno JB, Buglino JA, Boulton S, Chen H, Eswar N, He G, Huang R, Ilyin V, McMahan L, Pieper U, Ray S, Vidal M, Wang LK. Structural genomics: a pipeline for providing structures for the biologist. *Protein Sci.* 2002; 11:723–738. [PubMed: 11910018]
3. Fox BC, Gouliding C, Malkowski MG, Stewart L, Deacon A. Structural genomics: from genes to structures with valuable materials and many questions in between. *Nat. Methods.* 2008; 5:129–132. [PubMed: 18235432]
4. Dessailly BH, Nair R, Jaroszewski L, Fajardo JE, Kouranov A, Lee D, Fiser A, Godzik A, Rost B, Orengo C. PSI-2: structural genomics to cover protein domain family space. *Structure.* 2009; 17:869–881. [PubMed: 19523904]
5. Chandonia JM, Brenner SE. The impact of structural genomics: expectations and outcomes. *Science.* 2006; 311:347–351. [PubMed: 16424331]
6. Dunker AK, Brown CJ, Lawson JD, Iakoucheva LM, Obradovic Z. Intrinsic disorder and protein function. *Biochemistry.* 2002; 41:6573–6582. [PubMed: 12022860]
7. Dunker AK, Lawson JD, Brown CJ, Williams RM, Romero P, Oh JS, Oldfield CJ, Campen AM, Ratliff CM, Higgs KW, Ausio J, Nissen MS, Reeves R, Kang C, Kissinger CR, Bailey RW, Griswold MD, Chiu W, Garner EC, Obradovic Z. Intrinsically disordered protein. *J. Mol. Graph. Model.* 2001; 19:26–59. [PubMed: 11381529]
8. Dunker AK, Obradovic Z, Romero P, Garner EC, Brown CJ. Intrinsic protein disorder in complete genomes. *Genome Inform. Ser. Workshop Genome Inform.* 2000; 11:161–171.
9. Romero P, Obradovic Z, Li X, Garner EC, Brown CJ, Dunker AK. Sequence complexity of disordered protein. *Proteins.* 2001; 42:38–48. [PubMed: 11093259]
10. Ward JJ, Sodhi JS, McGuffin LJ, Buxton BF, Jones DT. Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. *J. Mol. Biol.* 2004; 337:635–645. [PubMed: 15019783]
11. Uversky VN. The mysterious unfoldome: structureless, underappreciated, yet vital part of any given proteome. *J. Biomed. Biotechnol.* 2010; 2010:568068. [PubMed: 20011072]
12. Xue B, Dunker AK, Uversky VN. Orderly order in protein intrinsic disorder distribution: disorder in 3500 proteomes from viruses and the three domains of life. *J. Biomol. Struct. Dyn.* 2012; 30:137–149. [PubMed: 22702725]
13. Ferron F, Longhi S, Canard B, Karlin D. A practical overview of protein disorder prediction methods. *Proteins.* 2006; 65:1–14. [PubMed: 16856179]
14. Esnouf RM, Hamer R, Sussman JL, Silman I, Trudgian D, Yang ZR, Prilusky J. Honing the in silico toolkit for detecting protein disorder. *Acta Crystallogr. D Biol. Crystallogr.* 2006; 62:1260–1266. [PubMed: 17001103]
15. Bourhis JM, Canard B, Longhi S. Predicting protein disorder and induced folding: from theoretical principles to practical applications. *Curr. Protein Pept. Sci.* 2007; 8:135–149. [PubMed: 17430195]
16. Dosztanyi Z, Sandor M, Tompa P, Simon I. Prediction of protein disorder at the domain level. *Curr. Protein Pept. Sci.* 2007; 8:161–171. [PubMed: 17430197]
17. Dosztanyi Z, Tompa P. Prediction of protein disorder. *Methods Mol. Biol.* 2008; 426:103–115. [PubMed: 18542859]
18. He B, Wang K, Liu Y, Xue B, Uversky VN, Dunker AK. Predicting intrinsic disorder in proteins: an overview. *Cell Res.* 2009; 19:929–949. [PubMed: 19597536]
19. Reeves R, Nissen MS. Purification and assays for high mobility group HMG-I(Y) protein function. *Methods Enzymol.* 1999; 304:155–188. [PubMed: 10372360]
20. Stewart AA, Ingebritsen TS, Cohen P. The protein phosphatases involved in cellular regulation. 5, Purification and properties of a Ca²⁺/calmodulin-dependent protein phosphatase (2B) from rabbit skeletal muscle. *Eur. J. Biochem.* 1983; 132:289–295. [PubMed: 6301828]
21. Petros AM, Medek A, Nettesheim DG, Kim DH, Yoon HS, Swift K, Matayoshi ED, Oltersdorf T, Fesik SW. Solution structure of the antiapoptotic protein bcl-2. *Proc. Natl. Acad. Sci. U. S. A.* 2001; 98:3012–3017. [PubMed: 11248023]
22. Bandaru V, Cooper W, Wallace SS, Doublet S. Overproduction, crystallization and preliminary crystallographic analysis of a novel human DNA-repair enzyme that recognizes oxidative DNA damage. *Acta Crystallogr. D Biol. Crystallogr.* 2004; 60:1142–1144. [PubMed: 15159582]

23. Bailey RW, Dunker AK, Brown CJ, Garner EC, Griswold MD. Clusterin, a binding protein with a molten globule-like region. *Biochemistry*. 2001; 40:11828–11840. [PubMed: 11570883]
24. Iakoucheva LM, Kimzey AL, Masselon CD, Bruce JE, Garner EC, Brown CJ, Dunker AK, Smith RD, Ackerman EJ. Identification of intrinsic order and disorder in the DNA repair protein XPA. *Protein Sci*. 2001; 10:560–571. [PubMed: 11344324]
25. Harauz G, Ishiyama N, Hill CM, Bates IR, Libich DS, Fares C. Myelin basic protein-diverse conformational states of an intrinsically unstructured protein and its roles in myelin assembly and multiple sclerosis. *Micron*. 2004; 35:503–542. [PubMed: 15219899]
26. Cary PD, King DS, Crane-Robinson C, Bradbury EM, Rabbani A, Goodwin GH, Johns EW. Structural studies on two high-mobility-group proteins from calf thymus, HMG-14 and HMG-20 (ubiquitin), and their interaction with DNA. *Eur. J. Biochem*. 1980; 112:577–580. [PubMed: 6257511]
27. Daughdrill GW, Chadsey MS, Karlinsey JE, Hughes KT, Dahlquist FW. The C-terminal half of the anti-sigma factor, FlgM becomes structured when bound to its target, sigma 28. *Nat. Struct. Biol*. 1997; 4:285–291. [PubMed: 9095196]
28. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The Protein Data Bank. *Nucleic Acids Res*. 2000; 28:235–242. [PubMed: 10592235]
29. Jurnak F. Structure of the GDP domain of EF-Tu and location of the amino acids homologous to ras oncogene proteins. *Science*. 1985; 230:32–36. [PubMed: 3898365]
30. Hobohm U, Scharf M, Schneider R, Sander C. Selection of representative protein data sets. *Protein Sci*. 1992; 1:409–417. [PubMed: 1304348]
31. Chen L, Oughtred R, Berman HM, Westbrook J. TargetDB: a target registration database for structural genomics projects. *Bioinformatics*. 2004; 20:2860–2862. [PubMed: 15130928]
32. Li X, Obradovic Z, Brown CJ, Garner EC, Dunker AK. Comparing predictors of disordered protein. *Genome Inform. Ser. Workshop Genome Inform*. 2000; 11:172–184.
33. Li X, Romero P, Rani M, Dunker AK, Obradovic Z. Predicting protein disorder for N-, C-, and internal regions. *Genome Inform. Ser. Workshop Genome Inform*. 1999; 10:30–40.
34. Romero P, Obradovic Z, Dunker AK. Sequence data analysis for long disordered regions prediction in the calcineurin family. *Genome Inform. Ser. Workshop Genome Inform*. 1997; 8:110–124.
35. Garner E, Cannon P, Romero P, Obradovic Z, Dunker AK. Predicting disordered regions from amino acid sequence: common themes despite differing structural characterization. *Genome Inform. Ser. Workshop Genome Inform*. 1998; 9:201–213.
36. Romero P, Obradovic Z, Kissinger CR, Villafranca JE, Dunker AK. Identifying disordered regions in proteins from amino acid sequences. *IEEE Int. Conf. Neural Netw*. 1997; 1:90–95.
37. Vucetic S, Radivojac P, Brown CJ, Dunker AK, Obradovic Z. Methods for improving protein disorder prediction. *IEEE Int. Conf. Neural Netw*. 2001; 4:2718–2723.
38. Peng K, Radivojac P, Vucetic S, Dunker AK, Obradovic Z. Length-dependent prediction of protein intrinsic disorder. *BMC Bioinformatics*. 2006; 7:208. [PubMed: 16618368]
39. Xue B, Dunbrack RL, Williams RW, Dunker AK, Uversky VN. PONDR-FIT: a meta-predictor of intrinsically disordered amino acids. *Biochim. Biophys. Acta*. 2010; 1804:996–1010. [PubMed: 20100603]
40. Peng K, Vucetic S, Radivojac P, Brown CJ, Dunker AK, Obradovic Z. Optimizing long intrinsic disorder predictors with protein evolutionary information. *J. Bioinform. Comput. Biol*. 2005; 3:35–50. [PubMed: 15751111]
41. Sing T, Sander O, Beerenwinkel N, Lengauer T. ROCr: visualizing classifier performance in R. *Bioinformatics*. 2005; 21:3940–3941. [PubMed: 16096348]
42. Santner AA, Croy CH, Vasanwala FH, Uversky VN, Van YY, Dunker AK. Sweeping away protein aggregation with entropic bristles: intrinsically disordered protein fusions enhance soluble expression. *Biochemistry*. 2012; 51:7250–7262. [PubMed: 22924672]
43. Fontana A, de Laureto PP, de Filippis V, Scaramella E, Zamboni M. Probing the partly folded states of proteins by limited proteolysis. *Fold. Des*. 1997; 2:R17–R26. [PubMed: 9135978]

44. Fontana A, Fassina G, Vita C, Dalzoppo D, Zamai M, Zambonin M. Correlation between sites of limited proteolysis and segmental mobility in thermolysin. *Biochemistry*. 1986; 25:1847–1851. [PubMed: 3707915]
45. Fontana A, Zambonin M, Polverino de Laureto P, De Filippis V, Clementi A, Scaramella E. Probing the conformational state of apomyoglobin by limited proteolysis. *J. Mol. Biol.* 1997; 266:223–230. [PubMed: 9047359]
46. Spolaore B, Bermejo R, Zambonin M, Fontana A. Protein interactions leading to conformational changes monitored by limited proteolysis: apo form and fragments of horse cytochrome c. *Biochemistry*. 2001; 40:9460–9468. [PubMed: 11583145]
47. Receveur-Brechot V, Bourhis JM, Uversky VN, Canard B, Longhi S. Assessing protein disorder and induced folding. *Proteins*. 2006; 62:24–45. [PubMed: 16287116]
48. Sedzik J, Kirschner DA. Is myelin basic protein crystallizable? *Neurochem. Res.* 1992; 17:157–166. [PubMed: 1371603]
49. Harauz G, Ladizhansky V, Boggs JM. Structural polymorphism and multifunctionality of myelin basic protein. *Biochemistry*. 2009; 48:8094–8104. [PubMed: 19642704]
50. Corollo D, Blair-Johnson M, Conrad J, Fiedler T, Sun D, Wang L, Ofengand J, Fenna R. Crystallization and characterization of a fragment of pseudouridine synthase RluC from *Escherichia coli*. *Acta Crystallogr. D Biol. Crystallogr.* 1999; 55:302–304. [PubMed: 10089432]
51. Orth P, Jekow P, Alonso JC, Hinrichs W. Proteolytic cleavage of gram-positive beta recombinaase is required for crystallization. *Protein Eng.* 1999; 12:371–373. [PubMed: 10360976]
52. Bracken C. NMR spin relaxation methods for characterization of disorder and folding in proteins. *J. Mol. Graph. Model.* 2001; 19:3–12. [PubMed: 11381530]
53. Metallo SJ. Intrinsically disordered proteins are potential drug targets. *Curr. Opin. Chem. Biol.* 2010; 14:481–488. [PubMed: 20598937]
54. Hammoudeh DJ, Follis AV, Prochownik EV, Metallo SJ. Multiple independent binding sites for small-molecule inhibitors on the oncoprotein c-Myc. *J. Am. Chem. Soc.* 2009; 131:7390–7401. [PubMed: 19432426]
55. Hurley TD, Yang J, Zhang L, Goodwin KD, Zou Q, Cortese M, Dunker AK, DePaoli-Roach AA. Structural basis for regulation of protein phosphatase 1 by inhibitors-2. *J. Biol. Chem.* 2007; 282:28874–28883. [PubMed: 17636256]
56. Huang HB, Chen YC, Tsai LH, Wang H, Lin FM, Horiuchi A, Greengard P, Nairn AC, Shiao MS, Lin TH. Backbone ¹H, ¹⁵N, and ¹³C resonance assignments of inhibitor-2—a protein inhibitor of protein phosphatase-1. *J. Biomol. NMR.* 2000; 17:359–360. [PubMed: 11014604]
57. Marsh JA, Dancheck B, Ragusa MJ, Allaire M, Forman-Kay JD, Peti W. Structural diversity in free and bound states of intrinsically disordered protein phosphatase 1 regulators. *Structure*. 2010; 18:1094–1103. [PubMed: 20826336]
58. Sickmeier M, Hamilton JA, LeGall T, Vacic V, Cortese MS, Tantos A, Szabo B, Tompa P, Chen J, Uversky VN, Obradovic Z, Dunker AK. DisProt: the database of disordered proteins. *Nucleic Acids Res.* 2007; 35:D786–D793. [PubMed: 17145717]
59. Uversky, VN.; Longhi, S. Instrumental analysis of intrinsically disordered proteins; assessing structure and conformation. In: Uversky, VN., editor. *The Wiley Series in Protein and Peptide Science*. John Wiley & Sons, Inc.; Hoboken, New Jersey, USA: 2010. p. 744
60. Uversky VN, Dunker AK. Multiparametric analysis of intrinsically disordered proteins: looking at intrinsic disorder through compound eyes. *Anal. Chem.* 2012; 84:2096–2104. [PubMed: 22242801]
61. Uversky, VN.; Dunker, AK. Experimental tools for the analysis of intrinsically disordered protein: volume I. In: Walker, J., editor. *Methods in Molecular Biology*. Humana Press; Totowa, NJ, USA: 2012.
62. Uversky, VN.; Dunker, AK. Experimental tools for the analysis of intrinsically disordered protein: volume II. In: Walker, J., editor. *Methods in Molecular Biology*. Humana Press; Totowa, NJ, USA: 2012.
63. Iakoucheva LM, Brown CJ, Lawson JD, Obradovic Z, Dunker AK. Intrinsic disorder in cell-signaling and cancer-associated proteins. *J. Mol. Biol.* 2002; 323:573–584. [PubMed: 12381310]

64. Dunker AK, Cortese MS, Romero P, Iakoucheva LM, Uversky VN. Flexible nets. The roles of intrinsic disorder in protein interaction networks. *FEBS J.* 2005; 272:5129–5148. [PubMed: 16218947]
65. Dyson HJ, Wright PE. Intrinsically unstructured proteins and their functions. *Nat. Rev. Mol. Cell Biol.* 2005; 6:197–208. [PubMed: 15738986]
66. Tompa P. Intrinsically unstructured proteins. *Trends Biochem. Sci.* 2002; 27:527–533. [PubMed: 12368089]
67. Uversky VN, Oldfield CJ, Dunker AK. Showing your ID: intrinsic disorder as an ID for recognition, regulation and cell signaling. *J. Mol. Recognit.* 2005; 18:343–384. [PubMed: 16094605]
68. Wright PE, Dyson HJ. Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm. *J. Mol. Biol.* 1999; 293:321–331. [PubMed: 10550212]
69. Zhu H, Snyder M. Protein arrays and microarrays. *Curr. Opin. Chem. Biol.* 2001; 5:40–45. [PubMed: 11166646]
70. Yang Z, Kollman JM, Pandi L, Doolittle RF. Crystal structure of native chicken fibrinogen at 2.7 Å resolution. *Biochemistry.* 2001; 40:12515–12523. [PubMed: 11601975]
71. Jiang X, Gurel O, Mendiaz EA, Stearns GW, Clogston CL, Lu HS, Osslund TD, Syed RS, Langley KE, Hendrickson WA. Structure of the active core of human stem cell factor and analysis of binding to its receptor kit. *EMBO J.* 2000; 19:3192–3203. [PubMed: 10880433]
72. Daumke O, Weyand M, Chakrabarti PP, Vetter IR, Wittinghofer A. The GTPase-activating protein Rap1GAP uses a catalytic asparagine. *Nature.* 2004; 429:197–201. [PubMed: 15141215]
73. Shaw N, Zhao M, Cheng C, Xu H, Saarikettu J, Li Y, Da Y, Yao Z, Silvennoinen O, Yang J, Liu ZJ, Wang BC, Rao Z. The multifunctional human p100 protein ‘hooks’ methylated ligands. *Nat. Struct. Mol. Biol.* 2007; 14:779–784. [PubMed: 17632523]
74. Deniziak M, Sauter C, Becker HD, Paulus CA, Giege R, Kern D. Deinococcus glutaminyl-tRNA synthetase is a chimera between proteins from an ancient and the modern pathways of aminoacyl-tRNA formation. *Nucleic Acids Res.* 2007; 35:1421–1431. [PubMed: 17284460]
75. Igonet S, Vulliez-Le Normand B, Faure G, Riottot MM, Kocken CH, Thomas AW, Bentley GA. Cross-reactivity studies of an anti-Plasmodium vivax apical membrane antigen I monoclonal antibody: binding and structural characterisation. *J. Mol. Biol.* 2007; 366:1523–1537. [PubMed: 17229439]
76. Macheboeuf P, Fischer DS, Brown T Jr. Zervosen A, Luxen A, Joris B, Dessen A, Schofield CJ. Structural and mechanistic basis of penicillin-binding protein inhibition by lactivicins. *Nat. Chem. Biol.* 2007; 3:565–569. [PubMed: 17676039]
77. Macheboeuf P, Di Guilmi AM, Job V, Vernet T, Dideberg O, Dessen A. Active site restructuring regulates ligand recognition in class A penicillin-binding proteins. *Proc. Natl. Acad. Sci. U. S. A.* 2005; 102:577–582. [PubMed: 15637155]
78. Healy S, McDonald MK, Wu X, Yue WW, Kochan G, Oppermann U, Gravel RA. Structural impact of human and *Escherichia coli* biotin carboxyl carrier proteins on biotin attachment. *Biochemistry.* 2010; 49:4687–4694. [PubMed: 20443544]
79. Mouilleron S, Badet-Denisot MA, Golinelli-Pimpaneau B. Ordering of C-terminal loop and glutaminase domains of glucosamine-6-phosphate synthase promotes sugar ring opening and formation of the ammonia channel. *J. Mol. Biol.* 2008; 377:1174–1185. [PubMed: 18295797]
80. Assenberg R, Delmas O, Ren J, Vidalain PO, Verma A, Larrous F, Graham SC, Tangy F, Grimes JM, Bourhy H. Structure of the nucleoprotein binding domain of Mokola virus phosphoprotein. *J. Virol.* 2010; 84:1089–1096. [PubMed: 19906936]
81. Chen G, Wang H, Robinson H, Cai J, Wan Y, Ke H. An insight into the pharmacophores of phosphodiesterase-5 inhibitors from synthetic and crystal structural studies. *Biochem. Pharmacol.* 2008; 75:1717–1728. [PubMed: 18346713]
82. Pei XY, Titman CM, Frank RA, Leeper FJ, Lulsi BF. Snapshots of catalysis in the E1 subunit of the pyruvate dehydrogenase multienzyme complex. *Structure.* 2008; 16:1860–1872. [PubMed: 19081062]
83. Wang H, Chumnarnsilpa S, Loonchanta A, Li Q, Kuan YM, Robine S, Larsson M, Mihalek I, Burtnick LD, Robinson RC. Helix straightening as an activation mechanism in the gelsolin

- superfamily of actin regulatory proteins. *J. Biol. Chem.* 2009; 284:21265–21269. [PubMed: 19491107]
84. Kollman JM, Pandi L, Sawaya MR, Riley M, Doolittle RF. Crystal structure of human fibrinogen. *Biochemistry.* 2009; 48:3877–3886. [PubMed: 19296670]
85. Lerch TF, Xie Q, Chapman MS. The structure of adeno-associated virus serotype 3B (AAV-3B): insights into receptor binding and immune evasion. *Virology.* 2010; 403:26–36. [PubMed: 20444480]
86. Lovering AL, Lin LY, Sewell EW, Spreter T, Brown ED, Strynadka NC. Structure of the bacterial teichoic acid polymerase TagF provides insights into membrane association and catalysis. *Nat. Struct. Mol. Biol.* 2010; 17:582–589. [PubMed: 20400947]
87. Jacques DA, Langley DB, Kuramitsu S, Yokoyama S, Trewella J, Guss JM. The structure of TTHA0988 from *Thermus thermophilus*, a Kipl-KipA homologue incorrectly annotated as an allophanate hydrolase. *Acta Crystallogr. D Biol. Crystallogr.* 2011; 67:105–111. [PubMed: 21245531]
88. Chen F, Venugopal V, Murray B, Rudenko G. The structure of neurexin 1alpha reveals features promoting a role as synaptic organizer. *Structure.* 2011; 19:779–789. [PubMed: 21620716]
89. Perriches T, Singleton MR. Structure of yeast kinetochore Ndc10 DNA-binding domain reveals unexpected evolutionary relationship to tyrosine recombinases. *J. Biol. Chem.* 2012; 287:5173–5179. [PubMed: 22215672]
90. Baranova E, Fronzes R, Garcia-Pino A, Van Gerven N, Papapostolou D, Pehau-Arnaudet G, Pardon E, Steyaert J, Howorka S, Remaut H. SbsB structure and lattice reconstruction unveil Ca²⁺-triggered S-layer assembly. *Nature.* 2012; 487:119–122. [PubMed: 22722836]

Fig. 1.

Cumulative histograms of (A) missing density regions by length and (B) structures by length of the longest missing density region in any protein chain in the structure. For each, missing density regions were filtered by the fraction of residues predicted to be disordered: no filtering (threshold of 0.0), at least partially disordered (threshold of 0.3), mostly disordered (threshold of 0.7), and completely disordered (threshold of 1.0).

Fig. 2. Cumulative histograms of predicted disorder in the ordered, disordered, and X-ray with disorder datasets. Disorder is summarized for each protein in the set by three different methods: (A) the percentage of residues per-protein predicted to be disordered, (B) the longest consecutive disorder prediction, and (C) the number of residues within predicted regions of disorder of 20 residues or longer. The dotted and dashed lines in each plot represent aggressive and conservative cutoffs, respectively, and are discussed in the text.

Fig. 3. Means of the percentage of predicted disorder for targets in the TargetTrack database. Targets are grouped by the year targets were first selected. Error bars give the 95% confidence interval of the mean.

Fig. 4.

Cumulative histograms for targets in the TargetTrack database for (A and C) percentage of predicted disorder and (B and D) number of residues within predicted regions of disorder of 20 residues or longer. Plots for two selection date thresholds are shown: (A and B) 2004 or earlier and (C and D) 2007 or later.

Fig. 5. Cumulative histogram for targets in the TargetTrack database select in 2004 or earlier by the linear combination of percentage of predicted disorder and number of residues within predicted regions of disorder of 20 residues or longer (equation given in text).

Fig. 6.

PONDRL® VL-XT predictions on the protein RluC. PONDRL® VL-XT prediction output values plotted versus residue number, where the order–disorder threshold is indicated by the thin horizontal line at a score of 0.5 (A). The region of this protein that was removed by proteolysis during crystallization trials is indicated by the thick gray bar. In an alternative representation, PONDRL® VL-XT output scores are indicated in the upper bar of (B) in binary form where black bars (score > 0.5) indicate disorder and white bars (score < 0.5) indicate order, with the sequence location indicated by the distance from the left end of the bar. The region of the protein that was crystallized is shown by the lower bar of (B), coded white to indicate the region is ordered.

Fig. 7.

Comparison of PONDR® predictions and protein fragments having 3-D structure. Proteins were chosen as representative of accurate and inaccurate PONDR® VL-XT predictions. The PONDR® VL-XT prediction for each protein is represented by the upper bar, where black disorder and white indicates order. The lower bars represent determined structures with PDB accessions as labeled, where white indicates defined coordinates and black represents missing coordinates or divergent NMR structures. Some examples are for proteins in isolation (1COK, 1TSR, 1NRC, 2U1A, 1IMX, 1HCP, 1F3M A/B, 1F3M C/D, 1F7W), other structures are for long protein fragments in complexes with other proteins (mdm2 in 1YCR, 1I4E, 1QBQ) or nucleic acid (1JK1, 2RAM), some are for homooligomers (1TGK, 1G50, 1FBU, 1ICE), and some are for short protein fragments in complexes with other proteins (1FOS, p53 in 1YCR, 3SAK, 1DT7).

Fig. 8.

Intrinsic disorder-based interaction between c-Myc, Max and DNA. Both c-Myc and Max are intrinsically disordered in their unbound states (left side), but fold into coiled-coil dimer upon interaction with DNA (right side) (PDB ID: 1NKP). c-Myc, Max, and DNA are shown as red ribbon, pink ribbon, and blue surface, respectively.

Table 1

The most extreme cases of apparent missing density in the PDB in terms of the longest missing density region (LMDR) and the fraction of residues in missing density regions (FMDR).

Structure	Residues	LMDR	FMDR	Reason for extreme missing density	Reference
1MIJ	2744	273	0.28	Disordered	[70]
1SCF	1092	122	0.59	Misannotation	[71]
1SRQ	1364	208	0.25	Mobile domain	[72]
2HQX	492	116	0.63	Likely proteolysis	[73]
2HZ7	851	220	0.35	Mobile domain	[74]
2IG2	671	217	0.32	Likely mobile domain ^a	
2J4W	883	378	0.47	Likely mobile domain	[75]
2J5L	1019	446	0.54	Likely mobile domain	[75]
2JCH	720	217	0.36	Proteolysis	[76,77]
2JE5	1440	222	0.36	Proteolysis	[77]
2JKU	94	36	0.63	Proteolysis	[78]
2VF4/2VF5	608	242	0.41	Mobile domain	[79]
2WZL	303	196	0.65	Proteolysis	[80]
3BJC	878	537	0.65	Misannotation	[81]
3DUF/3DVA/3DV0	3632	263	0.22	Misannotation	[82]
3EVY	478	147	0.65	Proteolysis ^b	
3FG7	796	277	0.73	Proteolysis	[83]
3GHG	5768	362	0.33	Proteolysis and disordered	[84]
3JZY	510	385	0.76	Proteolysis ^b	
3KIC/3KIE	14720	216	0.29	Likely disordered ^a	[85]
3L7L/3L7M/3L7J/3L7I	2916	313	0.44	Misannotation	[86]
3NG9	7360	216	0.29	Likely disordered ^a	
3ORE	988	220	0.25	Likely mobile domain	[87]
3R05/3QCW	2048	230	0.23	Likely mobile domain	[88]
4ACO	956	419	0.53	Misannotation	[89]
4AQ1	2044	264	0.22	Likely mobile domain	[90]

^aClassification based in part on disorder prediction.

^bBased on authors note in the PDB record.

Table 2

PONDR VL-XT error rates as a function of number of consecutive prediction of disorder. Pre-residue error rates are calculated as the number of amino acids in predicted disordered region of the given length or longer, divided by the number of residues in the set of ordered proteins. Per-protein error rates are calculated as the number of proteins containing at least one predicted disordered region of the indicated length or longer, divided by the number of proteins in the set of ordered proteins. These calculations were made using only residues with defined coordinates from the “ordered proteins” set.

Minimum number of consecutive predictions of disorder	Per-residue error rate	Per-protein error rate
1	20%	98%
10	15%	82%
20	8%	46%
30	4%	29%
40	2%	8%
50	0.5%	2%
60	0.1%	0.5%

Table 3

Logistic regression classification performance for combinations of disorder prediction-based parameters, longest predicted disordered region (LDR), number of residues in predicted disordered regions longer than 20 residues, and fraction of residues predicted to be disordered, measured by the area under the curve (AUC).

Disorder parameter			AUC
LDR	NDR	FDR	
+	-	-	0.578 $\pm$ 0.020
-	+	-	0.584 $\pm$ 0.013
-	-	+	0.533 $\pm$ 0.023
+	+	-	0.583 $\pm$ 0.020
+	-	+	0.589 $\pm$ 0.017
-	+	+	0.595 $\pm$ 0.020
+	+	+	0.600 $\pm$ 0.020