Intrinsic Protein Disorder in Complete Genomes

A. Keith Dunker[1]
dunker@disorder.chem.wsu.edu

Zoran Obradovic[2]
zoran@joda.cis.temple.edu

Pedro Romero[2]
promero@ai.sri.com

Ethan C. Garner[1]
egarner@disorder.chem.wsu.edu

Celeste J. Brown[1]
celesteb@disorder.chem.wsu.edu

[1]School of Molecular Biosciences, Washington State University, Pullman, WA 99164-4660, USA
[2]School of Electrical Engineering and Computer Sciences, Washington State University, Pullman, WA 99164-2752

Abstract

Intrinsic protein disorder refers to segments or to whole proteins that fail to fold completely on their own. Here we predicted disorder on protein sequences from 34 genomes, including 22 bacteria, 7 archaea, and 5 eucaryotes. Predicted disordered segments ≥ 50, ≥ 40, and ≥ 30 in length were determined as well as proteins estimated to be wholly disordered. The five eucaryotes were separated from bacteria and archaea by having the highest percentages of sequences predicted to have disordered segments ≥ 50 in length: from 25% for Plasmodium to 41% for Drosophila. Estimates of wholly disordered proteins in the bacteria ranged from 1% to 8%, averaging to 3±2%, estimates in various archaea ranged from 2 to 11%, plus an apparently anomalous 18%, averaging to 7±5% that drops to 5±3% if the high value is discarded. Estimates in the 5 eucarya ranged from 3 to 17%. The putative wholly disordered proteins were often ribosomal proteins, but in addition about equal numbers were of known and unknown function. Overall, intrinsic disorder appears to be a common, with eucaryotes perhaps having a higher percentage of native disorder than archaea or bacteria.

Keywords: intrinsic disorder, prediction, structural genomics

1 Introduction

A major effort in Bioinformatics is the prediction of function from amino acid sequence, with 3D structure viewed as a prerequisite for function [26]. Thus, associating a sequence with a particular structural family [25] or with a particular sequence motif [17] provides an avenue to predict function. One difficult of this {Sequence} → {3D Structure} → {Function} paradigm is the identification of distantly related sequences [9]. A second difficulty is that motifs such as the TIM barrel have evolved different functions [25]. Thus, knowing the structure gives not one but a set of likely functions. A third difficulty is that one protein can have two or more completely unrelated functions, known as *moonlighting* [20].

A more fundamental problem with the {Sequence} → {3D Structure} → {Function} paradigm is that it is simply not true for many proteins. Intrinsic disorder, not fixed 3D structure, is sometimes required for function [28, 36]. By intrinsic disorder we are referring to ensembles of structures, such as a random coil or molten globule, with the various members of the ensemble in equilibrium with each other.

From observations that variously shaped molecules bind competitively to serum albumin [21], 50 years ago Karush suggested that binding depends on an ensemble of structures in equilibrium. By now many additional examples of functional disorder are known; these include DNA recognition [33], enzymatic activation through proteolytic digestion [6], control of protein lifetimes [22], transport of an unfolded chain through a small orifice [8], and structural uncoupling of two or more domains by flexible linkers [19]. Given these many examples, identification of intrinsic disorder should be useful for inferring function.

We are studying predictions of disorder from amino acid sequence [10, 15, 23, 29-31, 35]. Here we report the results of application of a predictor of disorder to the proteins from 34 genomes. The results show that disorder is a very common element of protein structure and indicate that eucaryotes may have a higher proportion of intrinsic protein disorder than bacteria or archaea.

2 Materials and Methods

2.1 Databases of intrinsically ordered and disordered protein segments

The disordered data underlying the predictor utilized proteins having at least 40 consecutive disordered residues, with 1149 disordered residues in all. The X-ray characterized proteins had the following PDB Ids: 2tbv, 2ts1, 1aui, 1bgw, 1elo, 1bcl, 1ati, and 1lbh. The NMR-characterized sequences can be found in SWISS PROTEIN [2] (prio_mouse, h5_chick, flgm_salty, regn_lambd, hsf_klula, hmgi_human) or PIR [5] (S50866).

The ordered database was constructed from randomly selected segments from NRL_3D [27], which contains ordered residues. An amount of order to balance the disorder was captured.

To measure the false positive prediction rate, a database of ordered protein segments was constructed from PDB_SELECT_25 [16], which was based on grouping PDB proteins into families having > 25% sequence identity. ORDERED_PDB_SELECT_25 (O_PDB_S25) was derived from PDB_SELECT_25 by removing residues lacking backbone coordinates.

2.2 Genomic Databases

The amino acid sequences for known and putative proteins were obtained for 34 complete or mostly complete genomes from the NCBI (http://www.ncbi.nlm.nih.gov/PMGifs/Genomes/micr.html).

2.3 Predictor of intrinsic order and disorder

The predictor of natural protein disorder (PONDR) used for these studies represents a merger of 3 predictors: one for variously characterized long (VL1) regions of internal disorder and two for X-ray-characterized disorder located at the chain termini (XT), giving PONDR VL-XT. Each of these predictors is a simple neural network, with either 10-10-1 architecture (for VL1) or 8-8-1 architecture (for the two XTs). The inputs were attributes such as hydropathy or compositions of certain amino acids calculated as simple averages over windows of 21 residues for VL1 [31] and over windows of various lengths for the two XTs [23]. The two XT predictors are described in much more detail elsewhere [23], as is the VL predictor and its merger with the two XT predictors [31].

2.4 Application to genomic sequences and structural databases

PONDR VL-XT was applied to the protein sequences longer than 60 residues in length in 34 genomes. For *Methanococcus jannaschii*, *Escherichia coli,* and *Saccharomyces cerevisiae* the prediction results were parsed on the basis of having sequence similarity to a known structure through the use of PEDANT [12, 13]. The PEDANT matches were based on IMPALA searches [32] against a library of position-specific scoring matrices derived from each PDB sequence using BLAST [1].

3 Results

3.1 Predictor Error Estimation

To estimate false positive prediction rates, PONDR VL-XT was applied to O_PDB_S25. These are putative errors because some of the proteins in these two databases exist in the crystals as complexes and are disordered in the absence of their partners. These false positive error rates were determined by two methods of analysis: 1. the per-chain error and 2. the per-prediction error (Table 1). As we have shown before [30], as the length (L) of the disorder prediction increases, the false positive error rate drops rapidly; in this case from 1% of windows and 17% of chains with L ≥ 30 to 0.1% of windows and 2% of chains with L ≥ 50.

Table 1: False positive error rate for disorder predictions

Analysis	Fraction with L ≥ 50		Fraction with L ≥ 40		Fraction with L ≥ 30	
Per-chain error	17/1111	2%	69/1111	6%	189/1111	17%
Per-prediction error	173/166954	0.1%	702/177525	0.4%	2252/188479	1%

3.2 Prediction of wholly disordered proteins by cumulative distribution functions (CDFs)

A few proteins, such as FlgM [8], 4E-BP1 [11]; HMG-I(Y) [18], and neuromodulin [39], are disordered from end-to-end under physiological conditions, and yet they carry out function. Estimating the number of such wholly disordered proteins in the various genomes is of interest.

Figure 1 illustrates the method used for the identification of proteins that are likely to be wholly disordered. In this figure, 3 ways are shown for representing prediction data using 2 example proteins.

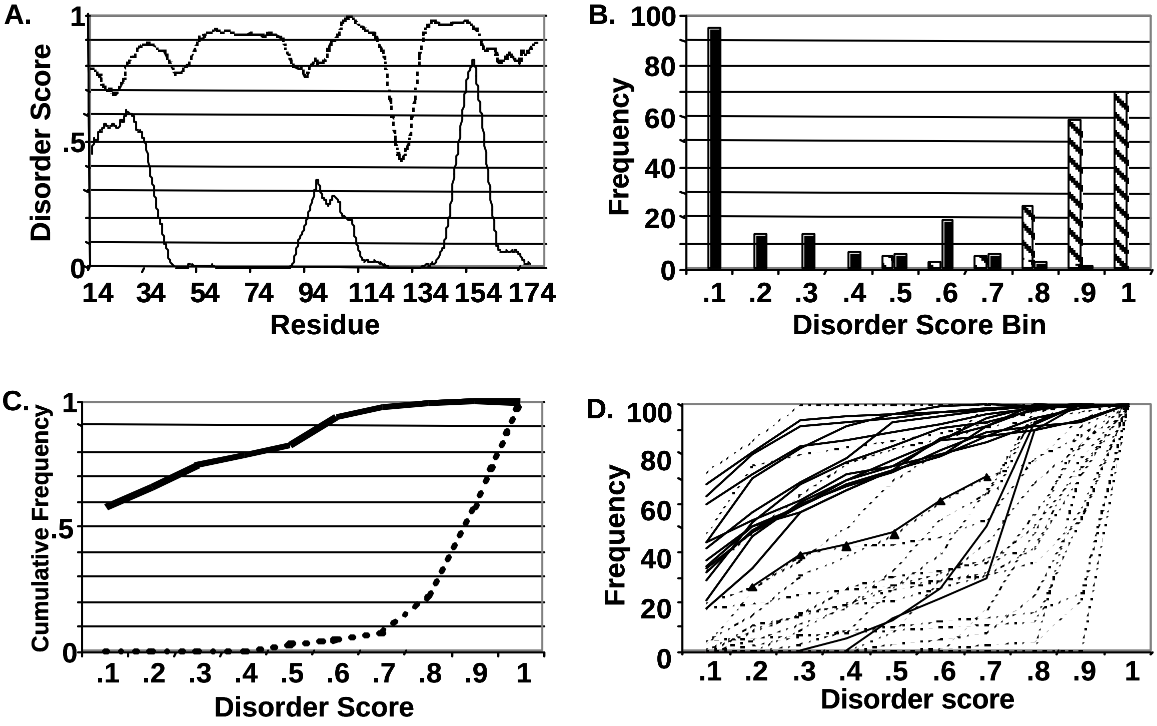

Figure 1. Development and application of CDF curves for identification of wholly disordered protein. A. Graph of disorder scores at each amino acid position for a disordered (dotted) and an ordered (bold) protein. B. Distribution of disorder scores for a disordered (cross hatched) and an ordered (solid) protein. C. CDF curves of disorder scores for disordered (dotted), and ordered (bold) proteins. D. A collection of fully ordered (bold) and fully disordered (dotted) proteins with an optimized boundary (▲).

The output of PONDR VL-XT is < 0.5 for a residue predicted to be ordered and > 0.5 for a residue predicted to be disordered, so ordered and wholly disordered proteins tend to lie on either side of this boundary (Fig. 1A). Alternatively, the predictions can be displayed as histograms (Fig. 1B). From each histogram, a cumulative distribution function (CDF) [34] can be calculated by determining the fraction of the distribution that lies below a given value on the x-axis (Fig. 1C). CDFs have the advantage that overlapping histograms can become completely separated curves.

The optimal boundary between datasets of completely ordered and completely disordered proteins was found by minimizing Error = #incorrect(O)/total(O) + #incorrect(D)/total(D), where #incorrect(O) indicates the number of ordered points incorrectly classified as disordered and #incorrect(D) indicates the number of disordered points incorrectly classified as ordered. This minimization was applied on a bin-by-bin basis with a bin size of 0.1. The optimization improved if boundary values for bins 0.1, 0.8, and 0.9 were simply omitted, thus yielding the boundary shown (Fig. 1D); CDF curves from the ordered and disordered protein sets are shown. Note the misclassification of two ordered proteins as disordered and four disordered as ordered.

3.3 Predicted disorder in 34 genomes

PONDR VL-XT was applied to the known and putative protein sequences of length L ≥ 60 in 34 genomes. The numbers of predicted to-be-disordered segments with L ≥ 50, ≥ 40, and ≥ 30 were determined. In addition, the CDF analysis for the identification of putative wholly disordered proteins was also carried out. The results of these two analyses for the 34 genomes are given in Table 2.

Table 2. Prediction of disorder in 34 genomes

Kingdom	Species	# seqs	L ≥ 30		L ≥ 40		L ≥ 50		CDF*	
Archaea	*Methanococcus jannaschii*	1714	367	21%	155	9%	71	4%	26	2%
Archaea	*Pyrococcus horikoshii*	2062	660	32%	330	16%	164	8%	70	3%
Archaea	*Pyrococcus abyssi*	1764	641	9%	338	19%	157	9%	62	4%
Archaea	*Archaeoglobus fulgidus*	2402	867	36%	492	20%	244	10%	93	4%
Archaea	*Methanobacterium thermoautotrophicum*	1869	971	52%	643	34%	365	20%	140	7%
Archaea	*Halobacterium sp.* NRC-1	2057	1096	53%	724	35%	484	24%	233	11%
Archaea	*Aeropyrum pernix* K1	2694	1547	57%	1010	37%	637	24%	490	18%
Bacteria	*Ureaplasma urealyticum*	611	87	14%	44	7%	14	2%	9	1%
Bacteria	*Rickettsia prowazekii*	834	129	15%	54	6%	23	3%	5	1%
Bacteria	*Borrelia burgdorferi*	845	110	13%	57	7%	26	3%	14	2%
Bacteria	*Campylobacter jejuni*	2309	328	14%	148	6%	80	3%	21	1%
Bacteria	*Mycoplasma genitalium*	480	77	16%	39	8%	20	4%	10	2%
Bacteria	*Helicobacter pylori*	1532	280	18%	140	9%	69	5%	24	2%
Bacteria	*Aquifex aeolicus*	1522	482	32%	234	15%	94	6%	29	2%
Bacteria	*Haemophilus influenzae*	1708	456	27%	227	13%	126	7%	27	2%
Bacteria	*Bacillus subtilis*	4093	1214	30%	622	15%	323	8%	87	2%
Bacteria	*Escherichia coli*	4281	1396	33%	731	17%	363	8%	107	2%
Bacteria	*Vibrio cholerae*	3815	1160	30%	595	16%	333	9%	93	2%
Bacteria	*Mycoplasma pneumoniae*	675	160	24%	95	14%	60	9%	14	2%
Bacteria	*Xylella fastidiosa*	2761	858	31%	463	17%	246	9%	103	4%
Bacteria	*Thermotoga maritima*	1842	670	36%	340	18%	165	9%	53	3%
Bacteria	*Neisseria meningitidis MC58*	2015	653	32%	351	17%	190	9%	64	3%
Bacteria	*Chlamydia pneumoniae*	1052	351	33%	185	18%	100	10%	40	4%
Bacteria	*Synechocystis sp*	3167	1106	35%	624	20%	338	11%	104	3%
Bacteria	*Chlamydia trachomatis*	894	314	35%	169	19%	99	11%	42	5%
Bacteria	*Treponema pallidum*	1028	392	38%	222	22%	115	11%	37	4%
Bacteria	*Pseudomonas aeruginosa*	5562	2314	42%	1310	24%	702	13%	183	3%
Bacteria	*Mycobacterium tuberculosis*	3916	2004	51%	1219	31%	747	19%	293	7%
Bacteria	*Deinococcus radiodurans* chr 1	2580	1335	52%	864	33%	534	21%	212	8%
Eukaryota	*Plasmodium falciparum* chr II, III	422	203	48%	147	35%	107	25%	11	3%
Eukaryota	*Caenorhabditis elegans*	17049	8304	49%	6156	36%	4636	27%	1322	8%
Eukaryota	*Arabodiopsis thaliana*	7849	4465	57%	3206	41%	2248	29%	653	8%
Eukaryota	*Saccharomyces cerevisiae*	6264	3373	54%	2527	40%	1858	30%	356	6%
Eukaryota	*Drosophila melanogaster*	13885	8771	63%	7031	51%	5651	41%	2403	17%

*numbers and percentages of chains predicted to be wholly disordered by the CDF analysis of Figure 1.

The percentage estimates in Table 2 are uncorrected for false positive error rates. From the per-chain false positive error rates in Table 1, the percentage values for L ≥ 50 should be reduced by ~ 2%, the values for L ≥ 40 by ~ 6% and the values for L ≥ 30 by ~ 17%. These corrections are only approximate.

Wholly disordered proteins should not form crystals, so any such protein that has high sequence similarity to a protein in PDB would be a candidate for a prediction error. PEDANT [12, 13] was used to compare the putative wholly disordered proteins of three representative genomes with the proteins in PDB, one for each kingdom: *M. jannaschii* for the archaea, *E. coli* for the bacteria, and *S. cerevisiae* for the eucaryotes. Of the 26, 107, and 356 putative wholly disordered proteins in *M. jannaschii, E. coli,* and *S. cerevisiae,* respectively, 2 in *M. jannaschii,* 20 in *E. coli* and 56 in *S. cerevisiae* were associated with proteins in PDB.

Further analysis (Table 3) shows that these associations might not all relate to prediction errors. For example, sometimes fragments of proteins rather than whole proteins are crystallized. Also, many intrinsically disordered proteins become ordered upon association with partners; such proteins can appear in PDB as ordered because the complex, not the individual protein, forms crystals. Finally, proteins in PDB may contain segments of disorder that are associated with the putative wholly disordered proteins.

As indicated in Table 3, 1 of the 2 putative wholly disordered proteins in *M. jannaschii*, all but 1 of the 20 such proteins in *E. coli,* and all but 2 of the 56 such proteins in *S. cerevisiae* fall into one of the categories suggesting that these proteins might be intrinsically disordered despite their appearance in PDB. More work is needed to better define the correspondence between the putative wholly disordered proteins and the related proteins in PDB, but these comparisons show that the error rate could be much lower than that suggested by simple associations with proteins in PDB.

Table 3. Putative wholly disordered proteins with sequence similarity to proteins of known 3D structure.

Organism	*M. jannaschii*	*E. coli*	*S. cerevisiae*
Total number	2	20	56
Only fragments visible		4	11
Bound to DNA		1	30
Bound to Protein		2	6
Bound to other Ligands		3	2
Bound within Di- or Multimers	1	1	2
Contain visible regions of disorder		8	3
Unbound Monomers	1	1	2

3.4 Functions of putative wholly disordered proteins

With regard to genomic studies of intrinsic disorder, of prime importance is the relationship between intrinsic disorder and protein function. An initial step towards this goal is to determine the functions of the putative wholly disordered proteins. Thus, we searched the various databases to determine which of the putative wholly disordered proteins have a known function, which have a likely function, and which are unidentified reading frames (URFs) [9]. The results are compiled in Table 4 for the three representative genomes.

Table 4: Categorization of putative wholly disordered proteins from 3 genomes

	M. jannaschii	*E. coli*	*S. cerevisiae*
Ribosomal	15	10	29
Known Function	1	36	157
Likely structural motif	5	13	28
URF	5	48	142

Ribosomal proteins stood out in the CDF analysis. Proteins in the known function category were grouped by having been studied directly or by having a high sequence similarity to a protein of known function. Too many different functions are represented to give them all. Proteins in the likely structural motif category have identifiable structural characteristics such as a probable transmembrane segment, although no relationship to proteins of known function could be established. We conjecture that a protein with a likely structural motif is more likely to be a real protein than any URF with no such feature.

4 Discussion

4.1 Prediction error rates

The relationship between per-prediction and per-chain error rates is simple (Table 1). The error rate of 0.1% came from the fact that the ~ 220,000 amino acids of O_PDB_S25 gave about 22 predictions of disorder of 50 or longer (e.g. 22/220,000 = 0.1%). Since this database contained 1,111 sequences, the per-chain error rate was about 22/1,111 or about 2%, indicating that two or more disorder predictions/chain were rare.

Comparing the per-chain error rates of Table 1 with the disorder predictions in Table 2 indicates that most of the disorder predictions in the genomes are much higher than in O_PSB_S25. It is tempting to just subtract the error rates from Table 1 from the prediction rates of Table 2 to give "corrected" estimates. However, there are problems with such a simple approach. First, the per-chain error rate increases as the length of an ordered protein increases. Thus, the error rate for O_PDB_S25 could be subtracted from that of a given genome only if the length distributions in the two datasets matched, but in general such a match is unlikely. Second, the segments in O_PDB_S25 would have to have similar false positive prediction rates as the ordered segments in the given genome, but this is unlikely to be true. O_PDB_S25 contains representative chains from each family of the current set of crystallized structures, whereas genomes don't contain just one protein from each family, but rather have sets of similar proteins with variable numbers of elements in the different sets.

In addition, the false negative errors, i.e. predictions of order for a residue or segment that is disordered, are not evaluated here. Our previous work shows that there are different types, or flavors, of disorder [15, 31, 35]. These flavors arise because the amino acid compositions of regions of intrinsic disorder are so variable; for example, one disordered region could be very rich in serine and glycine while another could be rich in lysine and arginine. Because of an uncertain number of flavors, it is not a simple task to estimate false negative error rates.

4.2 Commonness of intrinsic disorder

Sequence databases such as SWISS PROTEIN and PIR contain proteins with significantly higher fractions of predictions of long regions of disorder as compared to the sequences in the PDB [30, 31]. These data suggested that nature is quite rich in disorder and furthermore that PDB is strongly biased against intrinsic disorder, probably because of the requirement for crystallization [29-31].

Sequence databases have their own biases. For example, the leucine zippers appear more often than their occurrence in nature. Thus, compared to the various sequence databases, characterization of disorder on a genome-by-genome basis provides a better way of estimating the commonness of intrinsic disorder.

The bacteria and archaea exhibit a surprisingly wide range of disorder, with $L \geq 50$ ranging from 2 to 24%, or with the fractions of putative wholly disordered proteins ranging from 1 to 18% (Table 2). There is no clear separation of the archaea and bacteria based on the amount of predicted disorder. Further experimentation and study are needed to determine whether these large disorder variations are real.

The five eucaryotes exhibit the greatest amount of disorder as measured by $L \geq 50$ as compared to any of the archaea or bacteria. That is, the five eucaryotes yield 25-41% of chains with predicted disorder of $L \geq 50$ as compared to 24% for the highest bacterium. Data on more genomes are needed to determine whether eucaryotes indeed have more disorder as suggested here, especially given the two archaea that have nearly as much predicted disorder as the the lower range for the eucaryotes.

Disorder enables complexes with low affinity coupled with high specificity [10, 33] and also facilitates the binding of one molecule to many partners [10, 21, 22, 36]. These two characteristics were discovered by trying to understand proteins that are involved in signalling pathways. Thus, the higher amount of disorder in the eucaryotes might relate to a greater need for control and regulation as compared to the archaea and bacteria. If so, the wide range of predicted disorder in the archaea and bacteria might relate to differences in usages of signalling and control in these organisms. If this speculation were true, the number of regulatory proteins and hence the number of proteins with intrinsic disorder might be expected to increase as the total number of proteins increases. However, this simple expectation is not borne out; there is no clear relationship between total numbers of proteins and fractions of proteins with putatively disordered regions (see Table 2). So for now, these differences are unexplained.

4.3 Wholly disordered proteins and intrinsically disordered domains

Many wholly unfolded proteins and intrinsically disordered segments fold into specific 3D structure upon formation of a complex with a partner. For example, in figure 1D, two fully ordered proteins are predicted to be wholly disordered by the CDF classification – both of these proteins are bound DNA in the crystals used to determine their structures. Many of the putative wholly disordered proteins were shown to be likely to have biological partners (Table 3); binding to such partners could stabilize one member of the disordred ensemble and thereby cause a disorder-to-order transition upon binding.

One of the putative wholly disordered proteins from *M. jannaschii* has sequence similarity to a protein found in PDB (Table 3). This similar protein, from *Methanothermus fervidus*, crystallizes as a long coiled-coil dimer. Such coiled-coil proteins are disordered as monomers and become ordered upon self-association.

Many ribosomal proteins (Table 4) are unfolded when isolated and become ordered when they associate with the RNA. PONDR VL-XT recognized these proteins as wholly disordered even though no ribosomal proteins were used in the training set. However, the training set did contain other nucleic acid binding domains. Like the ribosomal proteins, these nucleic acid binding domains are disordered in the absence of their partner and have a large imbalance of positive charges. An especially interested feature of the ribosomal proteins is that they typically have small globular regions and long, extended tails that idiosyncratically meander among the RNA strands, probably acting as a glue for the various subregions of the RNA [4]. Thus, even in the ordered state, these ribosomal proteins resemble one member of an ensemble of a disordered protein.

Another interesting example is provided by 4E-BP1, which has been characterized by NMR to be wholly disordered. In an apparent error, PONDR predicts order for a 20 amino acid segment in the middle of this protein. However, this same segment of 4E-BP1 converts from disorder-to-order upon binding to its biological partner, eIF4E [24]. Thus, predictions of order in a structurally characterized region of disorder might indicate a possible binding site [15].

Like several related proteins, neuromodulin is disordered [39] and contains an IQ motif that binds calmodulin at low, not high, calciuim concentrations. These proteins are postulated to serve as calmodulin storage proteins. Phosphorylation of a site adjacent to the IQ motif abolishes calmodulin binding [7].

With regard to disordered proteins and domains that become ordered upon binding, the examples given above suggest that a wide range of behaviors can exist. For many wholly disordered sequences, such as the ribosomal proteins, the proteins simply associate with partners and remain ordered throughout their lifetimes. For others, such as 4E-BP1, the order/disorder transitions are coupled to association/dissociation reactions that are required for function. For still others, such as the neuromodulin group, phosphorylation affects the binding constant. Thus, the values for the binding constants for interactions involving intrinsically disordered proteins can span a very wide range and can be regulated by processes such as phosphorylation.

A single protein protein in the absence of any partners can be wholly disordered for 2 reasons: 1. the lack of suitable long-range interactions; and 2. a sequence that locally favors the disordered state. Of course, these same possibilities apply to segments of disorder that are within proteins that also have ordered domains. As discussed below, it is difficult to distinguish between these two possibilities with present tools.

Separated, almost equal-sized segments of the thioredoxin chain are disordered, but combine and fold into the correct 3D structures when mixed together [37, 38], showing that lack of suitable long-range partners can lead to the disordered state. Each of these segments, as expected, is predicted to be mostly ordered by PONDR VL-XT.

Balasubramanian et al. [3] recently cloned and expressed a set 12 proteins having high sequence identity to proteins from *Mycoplasma genitalium*, using the related genes from either *E. coli* or *B. subtilis*. These 12 proteins were chosen for study from their lack of sequence similarity to any protein of known function. Of these 12 proteins, 4 exhibited CD spectra more like those of unfolded rather than folded proteins. As shown in Table 2, 10 proteins in *M. genitalium* were predicted to be wholly disordered. None of these 10 match any of the 4 characterized as disordered by CD. Indeed, of 9 are ribosomal proteins and 1 is a cytoskeleton protein that likely contains a large amount of disorder. Therefore as expected, application of PONDR VL-XT to the cloned and expressed proteins yielded predictions of mostly order, rather than disorder, for all 4 proteins characterized to be mostly disordered by CD spectral analysis. There are two obvious explanations. First, these could be simply prediction errors by PONDR because these proteins are of disorder flavors not recognized by this predictor. Notice in Figure 1D that several structurally characterized regions of disorder fall into the ordered category by the CDF classifier; perhaps these proteins are likwise of unrecognized flavors of disorder. Alternatively, the four *M. genitalium* proteins could be parts of multi-subunit complexes, and if so they would lack the specific long-range interactions needed to induce folding. We are attempting to develop a strategy to distinguish between these two possibilities using sequence information alone. The difficulty lies in recognizing an intrinsically disordered region if its flavor has never been seen before.

4.4 Implications for structural genomics

The examples of intrinsic disorder studied in the laboratory and the predictions of disorder presented here suggest that structural genomics [14] will be seriously incomplete if only the structured parts of proteins are considered. On the computational side, improving predictions of disorder represents a significant challenge. On the experimental side, finding functions for structurally characterized regions of intrinsic disorder will be no easy task. With regard to the latter, especially difficult will be the discovery of new functions.

Acknowledgements

Support from NSF-CSE-IIS-9711532 to Z.O. and A.K.D and from NIH 1R01 LM06916 to A.K.D. and Z.O. are gratefully acknowledged.

References

[1] Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J., Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, Nucleic Acids Res., 25: 3389-3402, 1997.

[2] Bairoch, A. and Apweiler, R., The SWISS-PROT protein sequence data bank and its new supplement TREMBL, Nucleic Acids Res., 24: 21-25, 1996.

[3] Balasubramanian, S., Schneider, T., Gerstein, M., and Regan, L., Proteomics of *Mycoplasma genitalium*: identification and characterization of unannotated and atypical in a small model genome, Nucleic Acids Res., 28: 3075-3082, 2000.

[4] Ban, N., Nissen, P., Hansen, J., Moore, P.B., and Steitz, T.A. The complete atomic structure of the large ribosomal subunit at 2.4 A resolution. Science, 289: 905-20, 2000.

[5] Barker, W.C., Garavelli, J.S., Haft, D.H., Hunt, L.T., Marzec, C.R., Orcutt, B.C., Srinivasarao, G.Y., Yeh, L.S.L., Ledley, R.S., Mewes, H.W., Pfeiffer, F., and Tsugita, A., The PIR-international protein sequence database, Nucleic Acids Res., 26: 27-32, 1998.

[6] Bode, W., Schwager, P., and Huber, R., The transition of bovine trypsinogen to a trypsin-like state upon strong ligand binding. The refined crystal structures of the bovine trypsinogen-pancreatic trypsin

inhibitor complex and of its ternary complex with Ile-Val at 1.9Å resolution, J. Mol. Biol., 118: 99-112, 1978.

[7] Chakravarthy, B., Morley, P., and Whitfield, J., Ca2+-calmodulin and protein kinase Cs: a hypothetical synthesis of their conflicting convergences on shared substrate domains, Trends. Neurosci., 22: 12-16, 1999.

[8] Daughdrill, G.W., Chadsey, M.S., Karlinsey, J.E., Hughes, K.T., and Dahlquist, F.W., The C-terminal half of the anti-sigma factor, FlgM, becomes structured when bound to its target, sigma 28, Nat. Struct. Biol., 4: 285-291, 1997.

[9] Doolittle, R.F., Of URFS and ORFS. Mill Valley, CA: University Science Books, 1986.

[10] Dunker, A.K., Garner, E., Guilliot, S., Romero, P., Albrecht, K., Hart, J., Obradovic, Z., Kissinger, C., and Villafranca, J.E., Protein disorder and the evolution of molecular recognition: theory, predictions and observations, Pacific Symp. Biocomputing, 3: 473-484, 1998.

[11] Fletcher, C.M. and Wagner, G., The interaction of eIF4E with 4E-BP1 is an induced fit to a completely disordered protein, Protein Sci., 7: 1639-1642, 1998.

[12] Frishman, D. and Mewes, H.-W., PEDANTic genome analysis., Trends in Genetics, 13: 416-417, 1997.

[13] Frishman, D., Albermann, K., Hani, J., Heumann, K., Metanomski, A., Zollner, A., Mewes, H. -W., Functional and structural genomics using PEDANT, Bioinformtics, in press.

[14] Gaasterland, T., Structural genomics: Bioinformatics in the driver's seat, Nat. Biotechnology, 16: 625-627, 1998.

[15] Garner, E., Romero, P., Dunker, A.K., Brown, C., and Obradovic, Z., Predicting binding regions within disordered proteins, Genome Informatics, 10: 41-50, 1999.

[16] Hobohm, U. and Sander, C., Enlarged representative set of protein structures, Prot. Sci., 3: 522-524, 1994.

[17] Hofmann, K., Bucher, P., Falquet, L., and Bairoch, A., The PROSITE database, its status in 1999, Nucleic Acids Res, 27: 215-219, 1999.

[18] Huth, J.R., Bewley, C.A., Nissen, M.S., Evans, J.N., Reeves, R., Gronenborn, A.M., and Clore, G.M., The solution structure of an HMG-I(Y)-DNA complex defines a new architectural minor groove binding motif, Nat. Struct. Biol., 4: 657-665, 1997.

[19] Jaffray, E., Wood, K.M., and Hay, R.T., Domain organization of I kappa B alpha and sites of interaction with NF- kappa B p65, Mol. Cell. Biol., 15: 2166-2172, 1995.

[20] Jeffery, C.J., Moonlighting proteins, Trends Biochem Sci, 24: 8-11, 1999.

[21] Karush, F., Heterogeneity of the binding sites of bovine serum albumin, JACS, 72: 2705-2713, 1950.

[22] Kriwacki, R.W., Hengst, L., Tennant, L., Reed, S.I., and Wright, P.E., Structural studies of p21Waf1/Cip1/Sdi1 in the free and Cdk2-bound state: conformational disorder mediates binding diversity, Proc. Natl. Acad. Sci. USA, 93: 11504-11509, 1996.

[23] Li, X., Romero, P., Rani, M., Dunker, A.K., and Obradovic, Z., Predicting protein disorder for N-, C-, and internal regions, Genome Informatics, 10: 30-40, 1999.

[24] Marcotrigiano, J., Gingras, A.C., Sonenberg, N., and Burley, S.K., Cap-dependent translation initiation in eukaryotes is regulated by a molecular mimic of eIF4G, Mol Cell, 3: 707-16, 1999.

[25] Orengo, C.A., Michie, A.D., Jones, S., Jones, D.T., Swindells, M.B., and Thornton, J.M., CATH--a hierarchic classification of protein domain structures, Structure, 5: 1093-108, 1997.

[26] Orengo, C.A., Pearl, F.M., Bray, J.E., Todd, A.E., Martin, A.C., Lo Conte, L., and Thornton, J.M., The CATH Database provides insights into protein structure/function relationships, Nucleic Acids Res, 27: 275-9, 1999.

[27] Pattabiraman, N., Namboodiri, K., Lowrey, A., and Gaber, B.P., NRL-3D: a sequence-structure database derived from the protein data bank (PDB) and searchable within the PIR environment, Protein Seq. Data Anal., 3: 387-405, 1990.

[28] Plaxco, K.W. and Gross, M., The importance of being unfolded, Nature, 386: 657, 659, 1997.

[29] Romero, P., Obradovic, Z., Kissinger, C.R., Villafranca, J.E., and Dunker, A.K., Identifying disordered regions in proteins from amino acid sequences, Proc. I.E.E.E. International Conference on Neural Networks, 1: 90-95, 1997.

[30] Romero, P., Obradovic, Z., Kissinger, C.R., Villafranca, J.E., Guilliot, S., Garner, E., and Dunker, A.K., Thousands of proteins likely to have long disordered regions, Pacific Symp. Biocomputing, 3: 437-448, 1998.

[31] Romero, P., Obradovic, Z., Li, X., Garner, E.C., Brown, C.J., and Dunker, A.K., Sequence Complexity of Disordered Protein, Proteins: Struc., Funct., Gen., 42:38-48, 2001.

[32] Schaffer, A. A., Wolf, Y. I., Ponting, C. P., Koonin, E. V., Aravind, L., Altschul, S. F., IMPALA: matching a protein sequence against a collection of PSI-BLAST-constructed position-specific score matrices. Bioinformatics, 15: 1000-1011, 1999.

[33] Schulz, G.E., Nucleotide Binding Proteins, in Molecular Mechanism of Biological Recognition, M. Balaban, Ed. New York: Elsevier/North-Holland Biomedical Press, 1979, 79-94.

[34] Sprent, P., Applied Nonparametric Statistical Methods: 2nd ed. London: Chapman and Hall, 1993.

[35] Wang, J., Family-specific Neural Network Predicters and Different Flavors of Disordered Regions, Master's Thesis, Washington State University, 2000

[36] Wright, P.E. and Dyson, H.J., Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm, J. Mol. Biol., 293: 321-331, 1999.

[37] Yang, X.M., Yu, W.F., Fuchs, J., Rizo, J., and Tasayco, M.L., NMR evidence for teh reassembly of an alpha/beta domain after cleavage of an alpha-helix: implications for protein design, J. Ame. Chem. Soc., 120: 7985-7986, 1998.

[38] Yang, X.-M., Georgescu, R.E., Li, J.-H., Yu, W.-F., Haierhan, and Tasayco, M.L., Recognition between disordered polypeptide chains from cleavage of an α/β domain: self- versus non-self-association, Pacific Symp. Biocomputing, 4: 590-600, 1999.

[39] Zhang, M., Vogel, H.J., and Zwiers, H., Nuclear magnetic resonance studies of the structure of B50/neuromodulin and its interaction with calmodulin, Biochem. Cell Biol., 72: 109-116, 1994.