doi:10.1016/S0022-2836(02)00969-5 available online at http://www.idealibrary.com on IDEAL® J. Mol. Biol. (2002) not known, 1–12

JMB

Intrinsic Disorder in Cell-signaling and Cancer-associated Proteins

Lilia M. Iakoucheva[1], Celeste J. Brown[1], J. David Lawson[1] Zoran Obradović[2] and A. Keith Dunker[1]*

[1]*Department of Biochemistry and Biophysics, School of Molecular Biosciences Washington State University Pullman, WA 99164-4660 USA*

[2]*Center for Information Science and Technology, Temple University, Philadelphia, PA 19122, USA*

**Corresponding author*

The number of intrinsically disordered proteins known to be involved in cell-signaling and regulation is growing rapidly. To test for a generalized involvement of intrinsic disorder in signaling and cancer, we applied a neural network predictor of natural disordered regions (PONDR VL-XT) to four protein datasets: human cancer-associated proteins (HCAP), signaling proteins (AfCS), eukaryotic proteins from SWISS-PROT (EU_SW) and non-homologous protein segments with well-defined (ordered) 3D structure (O_PDB_S25). PONDR VL-XT predicts ≥30 consecutive disordered residues for 79(±5)%, 66(±6)%, 47(±4)% and 13(±4)% of the proteins from HCAP, AfCS, EU_SW, and O_PDB_S25, respectively, indicating significantly more intrinsic disorder in cancer-associated and signaling proteins as compared to the two control sets. The disorder analysis was extended to 11 additional functionally diverse categories of human proteins from SWISS-PROT. The proteins involved in metabolism, biosynthesis, and degradation together with kinases, inhibitors, transport, G-protein coupled receptors, and membrane proteins are predicted to have at least twofold less disorder than regulatory, cancer-associated and cytoskeletal proteins. In contrast to 44.5% of the proteins from representative non-membrane categories, just 17.3% of the cancer-associated proteins had sequence alignments with structures in the Protein Data Bank covering at least 75% of their lengths. This relative lack of structural information correlated with the greater amount of predicted disorder in the HCAP dataset. A comparison of disorder predictions with the experimental structural data for a subset of the HCAP proteins indicated good agreement between prediction and observation. Our data suggest that intrinsically unstructured proteins play key roles in cell-signaling, regulation and cancer, where coupled folding and binding is a common mechanism.

Keywords: intrinsic disorder; unstructured protein; protein folding; cell signaling; cancer

Introduction

The dominating concept that protein structure determines protein function is undergoing re-evaluation. Interest in intrinsically unstructured proteins is rising because of recognition that biological function derives from ordered 3D structure and from the lack of specific structure. Indeed, some proteins require the absence of prior 3D structure to carry out their functions.[1–3] A literature review including more than 90 proteins revealed that a majority of known disordered proteins or domains were involved in cell-signaling or regulation *via* non-catalytic interactions with DNA, RNA, or other proteins.[4] Such unstructured regions become folded upon binding to their targets, thereby confirming that initial 3D structure is not required for biomolecular recognition.[5–7]

Abbreviations used: CDK, cyclin-dependent kinase; eIF4E, translation initiation factor (eIF) 4F; EU_SW, eukaryotic proteins from SWISS-PROT; HCAP, human cancer-associated proteins; PONDR, predictor of natural disordered regions; SAM domain, sterile alpha motif domain; TBP, TATA box-binding protein; TFE, trifluoroethanol; 4E-BP, 4E binding protein; 53Bp1, p53 binding protein 1; 53Bp2, p53 binding protein 2.

E-mail address of the corresponding author: dunker@disorder.chem.wsu.edu

0022-2836/02/$ - see front matter

Molecular recognition involving intrinsically disordered proteins has two features that provide important functional advantages for signaling and regulation. First, disordered regions can bind their targets with high specificity and low affinity.[2,8] Second, intrinsic disorder promotes binding diversity by enabling proteins to interact with numerous partners.[7,9] Thus, hubs and nodes in signaling networks are likely to include proteins with extended disordered regions. In support of this possibility, two well-studied proteins, p53 and HMGA, interact with their multiple partners mostly *via* regions of intrinsic disorder.[10]

A comparison of two complete eukaryotic genomes, a unicellular yeast *Saccharomyces cerevisiae* and multicellular nematode *Caenorhabditis elegans*, suggests that multicellular organisms have developed elaborate signal transduction and regulatory control by employing novel proteins. Many of these proteins re-use evolutionarily conserved domains whose functions were initially unrelated to signal transduction.[11] Flexbility and disorder in linkers connecting these domains in multidomain eukaryotic proteins appears to be an important characteristic of multicellularity.

The prevalence of intrinsically unstructured proteins in eukaryotic genomes in comparison to bacteria and archaea[12] together with numerous examples of unstructured regions in regulatory proteins[7,13] may reflect the greater need for disorder-associated signaling and regulation in nucleated cells.[10] Here, we apply a predictor of intrinsically unfolded protein regions to investigate disorder in cancer-associated and cell-signaling proteins. We then compare the amount of predicted disorder among diverse protein categories and correlate our predictions with the available structural information. The results support a general involvement of intrinsically unstructured proteins in cell-signaling, regulation and human cancer.

Results and Discussion

Disorder prediction on cancer-associated and cell-signaling proteins

To test for an association between signaling and intrinsic disorder, we used a predictor of natural disordered regions (PONDR VL-XT)[14] to systematically analyze the intrinsic disorder tendencies in four protein datasets (Table 1): (1) human cancer-associated proteins from SWISS-PROT (HCAP); (2) signaling proteins collected by the Alliance for Cellular Signaling (AfCS); (3) the eukaryotic proteins from SWISS-PROT (EU_SW); and (4). a set of non-homologous protein segments with well-defined (ordered) 3D structure from the Protein Data Bank Select 25 (O_PDB_S25). The O_PDB_S25 dataset provides a non-redundant control for estimating the false-positive disorder prediction error rate.

Table 1. Description of four protein datasets

Database name	No. proteins in database	No. proteins for predict.	Max. protein length (res.)	Average length (res.)	Median length (res.)
HCAP	231	231	3969	620	462
AfCS	2329	2325	5038	588	465
EU_SW	53,630	53,602	6669	408	334
PDB_S25	1138	1136	965	206	171

Proteins shorter than 30 amino acid residues (res.) were eliminated from the PONDR VL-XT predictions.

The analysis of PONDR VL-XT predictions demonstrates that predicted disorder followed the ranking HCAP > AfCS > EU_SW ≫ O_PDB_S25 (Figure 1). The same ranking was observed whether the results were presented as percentages of proteins (Figure 1(a)) or as percentages of residues (Figure 1(b)). The percentages of proteins (± two standard errors) with 30 or more consecutive residues predicted to be disordered were 79(±5)% for HCAP, 66(±6)% for AfCS, 47(±4)% for EU_SW, and 13(±4)% for O_PDB_S25, with the errors estimated as described in Materials and Methods. Thus, ~1.6-fold and ~1.4-fold more of the HCAP and AfCS proteins, respectively, had predicted disordered regions of ≥30 consecutive residues as compared to the EU_SW proteins, while ~3.6-fold more of the EU_SW proteins had such regions of predicted disorder in comparison to the O_PSB_S25 proteins. When analyzed by percentages of residues, the HCAP proteins had ~1.8-fold more predicted disorder than EU_SW, and ~8.6-fold more disorder than O_PDB_S25 for regions with ≥30 consecutive disorder predictions; these estimates of disorder rise to ~2.6-fold and to >350-fold, respectively, for regions with ≥60 consecutive disorder predictions (Figure 1(b)). Thus, HCAP and AfCS proteins were innately richer in predicted disorder than the typical eukaryotic proteins.

Signaling and cancer-associated proteins are highly interrelated,[15] and the increased amount of predicted disorder in these two protein datasets reflects this connection. Over-expression or constitutive activation of some oncogenes may contribute to the loss of cell-cycle control observed in many tumors.[16] A large number of proto-oncogenes (i.e. c-*jun*, c-*fos*, c-*myc*) code for transcription factors required for cell-cycle progression and cell differentiation. Experimental evidence of disorder in signaling and oncoproteins further support our disorder predictions. For example, the N terminus of tumor suppressor Arf regulates p53 function through binding to oncoprotein Hdm2 and is unstructured in solution.[17] The C-terminal activation domain of c-*fos* in its biologically active form is structurally disordered.[18] This domain interacts directly with multiple transcription factors: TBP, TFIIH, CBP and Smad3,[19–21] and activates transcription in different cellular processes. Disorder would be a significant factor contributing to the conformational freedom of this domain and

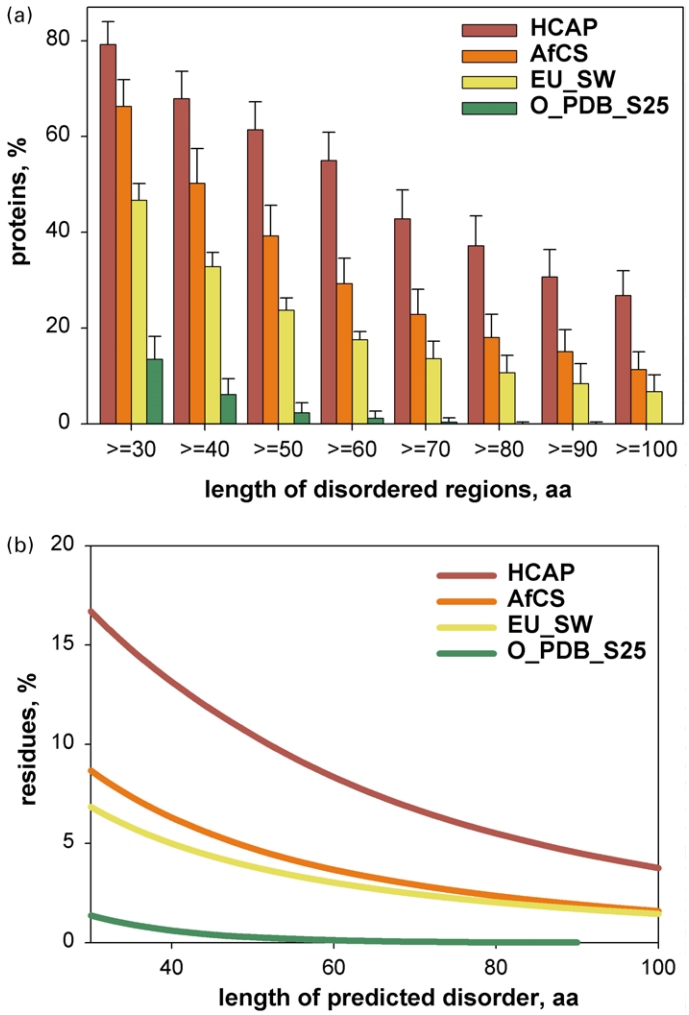

Figure 1. PONDR VL-XT disorder prediction results on four datasets. (a) Percentages of proteins in the four datasets with ≥30 to ≥100 consecutive residues predicted to be disordered. The error bars represent 95% confidence intervals and were calculated as described in Materials and Methods. The O_PDB_S25 dataset provides a mostly non-redundant control for estimating the false-positive disorder prediction error rate. (b) Percentages of residues in the four datasets predicted to be disordered within segments of length ≥ the value on the *x*-axis.

allowing it to associate with numerous partners. Two more examples of intrinsically disordered domains that become ordered upon synergistic folding are ACTR and CBP.[22]

Eukaryotic proteins often contain multiple structured domains connected by flexible linkers. Experimentation on a small collection of linkers indicated that high percentages of their residues were predicted to be disordered by PONDR VL-XT (our unpublished results). Thus, the common occurrence of multiple domains connected by flexible linkers probably underlies the finding that 47(±4)% of EU_SW have ≥30 consecutive residues predicted to be disordered. The signaling and cancer-associated proteins, however, are even richer in predicted disorder than typical eukaryotic proteins. This additional disorder is proposed to relate to the signaling and regulatory functions of these proteins. Such interpretation is supported by our analysis of the functions for about 90 proteins with long regions of disorder.[4]

Disorder analysis of distinct protein categories from SWISS-PROT

We expanded our disorder analysis to include 11 additional datasets representing different types of human proteins from SWISS-PROT (Table 2 and Figure 2). The comparison of mean protein lengths for each dataset shows that they vary over a range of about 30% with two exceptions: cytoskeletal and ribosomal proteins (Figure 2(a)). Cytoskeletal proteins are, in general, considerably longer, while ribosomal proteins are, on average, much shorter. The differences in sequence lengths between the datasets are important for our disorder analysis: the longer proteins would be expected by chance to have longer regions of predicted disorder.

PONDR VL-XT was applied to the different protein categories from SWISS-PROT (Figure 2(b) and (c)). Similar to our previous analysis, cancer-associated and regulatory proteins show significantly more disorder than most of the other protein categories, whether expressed as percentage of proteins (Figure 2(b)) or as percentage of residues (Figure 2(c)). A comparable amount of disorder is predicted for cytoskeletal proteins, and starting from ≥40 residues, the percentage of proteins with predicted disorder in these three protein categories is significantly higher (up to 2.5-fold) than in all other datasets.

The increased lengths of the cytoskeletal proteins (Table 2 and Figure 2(a)) may partially account for the higher percentage of proteins with predicted

Table 2. Description of 11 protein datasets from SWISS-PROT

Database name	No. proteins in database	No. proteins for predict.	Max. protein length (res.)	Average length (res.)	Median length (res.)
Regulation	851	851	3969	548	458
Cytoskeletal	134	134	6669	1044	732
Ribosomal	104	103	547	187	158
Membrane	179	179	3674	632	503
Transport	593	593	4563	545	468
Biosynthesis	245	245	2504	509	445
Inhibitors	113	113	4829	460	352
Kinases	95	95	3056	564	419
Metabolism	112	112	4563	618	511
Degradation	59	56	1290	519	469
G-pr.coup.receptors	339	339	1584	416	365

Proteins shorter than 30 amino acid residues (res.) were eliminated from the PONDR VL-XT predictions.

disordered regions in this dataset (Figure 2(b)). The analysis on a per residue basis (Figure 2(c)), however, also indicated a large amount of disorder, implying that the increased length does not completely explain the higher percentage of predicted disorder in this category. To test the possibility that disorder in the cytoskeletal set is linked to the coiled-coil regions, COILS predictions[23] were correlated with PONDR VL-XT disorder predictions. For the cancer-associated, regulatory, and cytoskeletal proteins, 4%, 2% and 10% of all residues, respectively, and 7.5%, 2.1%, and 16.6% of putatively disordered residues, respectively, were predicted to be in coiled-coil helices. These data suggest that cytoskeletal proteins are indeed richer in coiled coils as compared to the other two categories, and that coiled coils are often predicted to be disordered by PONDR VL-XT. However, coiled coils accounted for less than 20% of the putatively disordered residues in cytoskeletal proteins. An additional contributor to the high disorder content in this category is that most of these proteins form filamentous structures in addition to coiled-coil assemblies, and the formation of such protein–protein interactions often involves regions of intrinsic disorder.[24] A final comment on the grouping of cytoskeletal proteins with regulatory and cancer-associated proteins is that many cytoskeletal proteins are involved in cell-signaling,[25,26] playing key roles in T-cell activation,[27] platelet-signaling,[28] and cancer development.[29]

Nearly 68(±11)% of ribosomal proteins have predicted disordered regions of ≥30 residues (Figure 2(b)). Although this value is somewhat lower than the 85(±2)% of regulatory proteins and 79(±4)% of cancer-associated proteins, the disorder in ribosomal proteins should still be considered substantial due to the decreased average length of proteins from this category (Figure 2(a)). Moreover, for regions of ≤30 residues, the percentage of disordered residues in ribosomal dataset becomes comparable to that in HCAP, regulatory and cytoskeletal categories (data not shown). Our predictions are strengthened by the experimental evidence of disorder in numerous ribosomal proteins when separated from the ribosome.[30,31] The high ratio of charged to hydrophobic amino acid residues has been suggested as the likely cause of the observed disorder in these proteins.[32,33]

Although the regulatory (851 proteins) and AfCS (2329 proteins) datasets differ quantitatively and qualitatively, we observed similar disorder prediction results for both (compare Figures 1(a) and 2(b)). Our analysis applied to these two independently constructed sets strongly supports the increased amount of disorder in proteins involved in cell-signaling and regulation. Interestingly, the proteins that perform mainly catalytic cellular functions (for example, metabolism, biosynthesis, and degradation), have significantly less predicted disorder. We suggest that regulatory proteins or domains are disordered without their binding partners, whereas catalytic proteins or domains form well-defined, folded 3D structure even in the absence of their substrates. As discussed previously, molecular recognition by ordered protein structure may be involved predominantly in catalysis, while molecular recognition by disordered structure may be especially important for regulation and signaling.[4]

The 3D structural information for representative protein datasets

We previously observed that proteins in the PIR and SWISS-PROT databases contain substantially more predicted disorder than the proteins in PDB, evidently because the requirement for crystallization biases PDB against proteins with long regions of disorder.[34,35] If indeed HCAP proteins are as rich in disorder as predicted, these proteins should be under-represented in PDB. To test this possibility, we searched PDB for homologues using the gapped-BLAST alignment algorithm (Materials and Methods). In many cases, single sequences were homologous to multiple protein structures in PDB. The percentage of each sequence that aligned with one or more 3D structures (PDB coverage) was plotted *versus* the length of each protein in Figure 3. The plots were constructed for HCAP (Figure 3(a)) and three representative, non-membrane protein control sets: biosynthesis

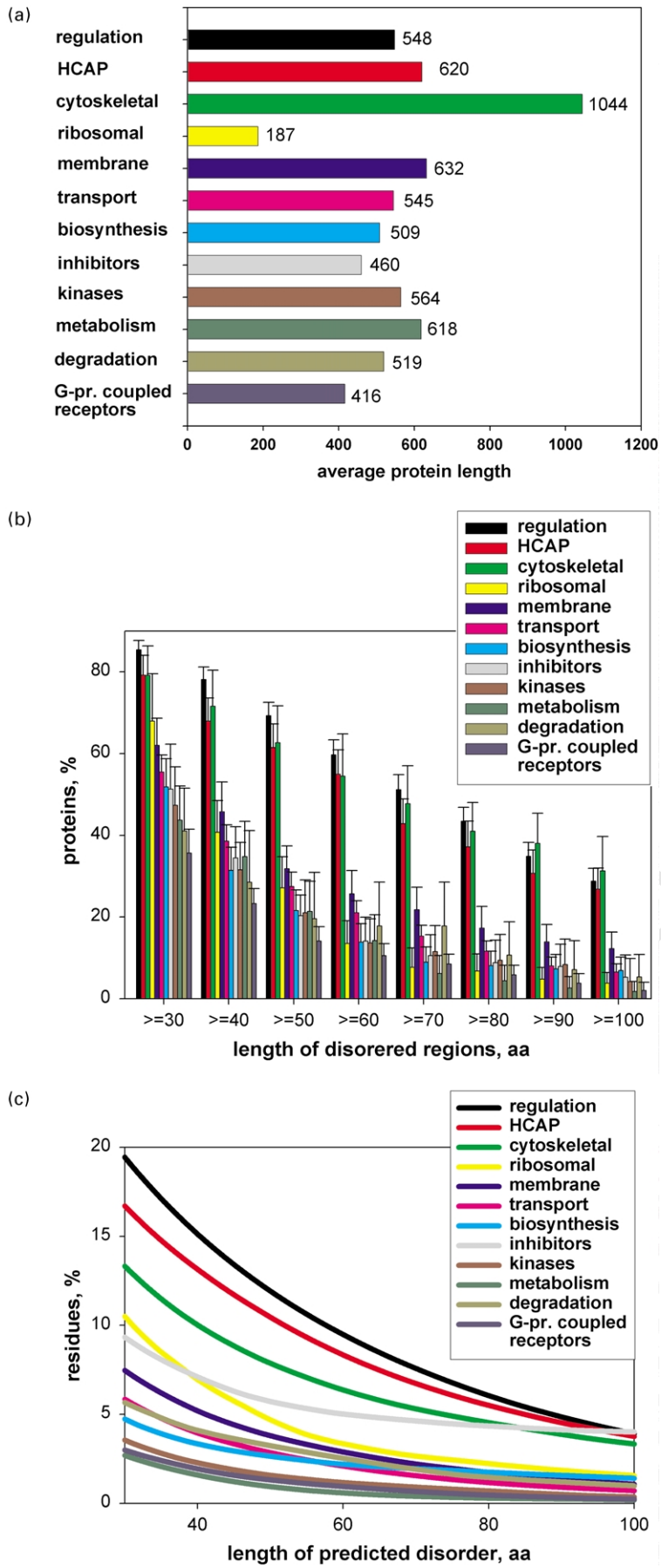

Figure 2. Disorder analysis of functional protein categories from SWISS-PROT. (a) Average length distribution of human proteins from 12 functional categories. The numbers indicate the average protein length for each dataset. (b) Predicted disorder in proteins from SWISS-PROT. The error bars represent 95% confidence intervals and were calculated as described in Materials and Methods. (c) Percentages of residues in the 12 datasets predicted to be disordered within segments of length ≥ the value on the *x*-axis.

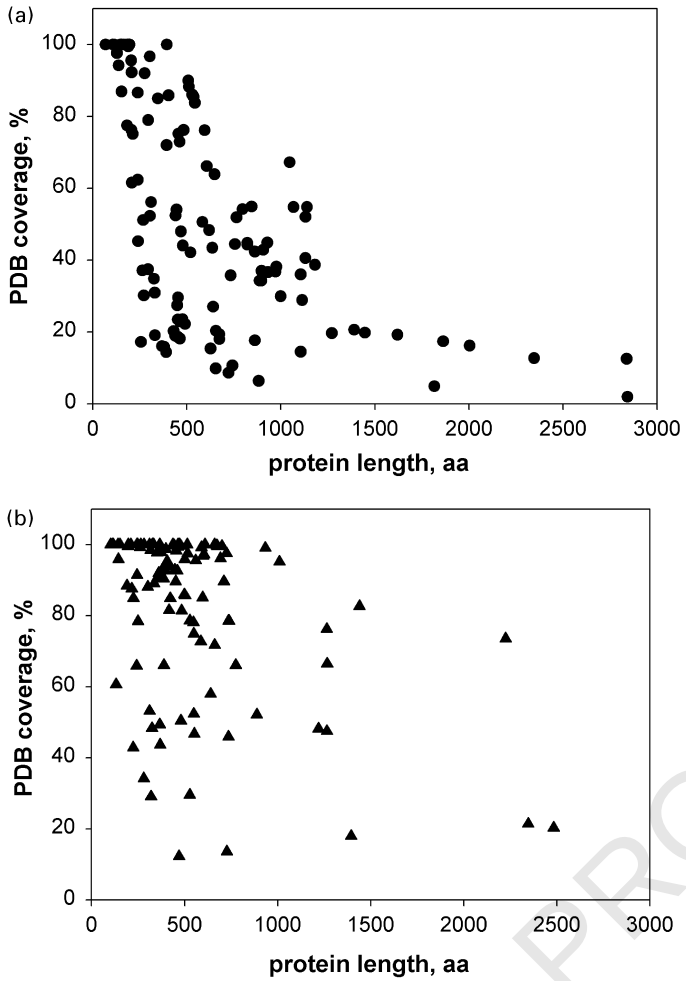

Figure 3 (*legend opposite*)

(Figure 3(b)), degradation (Figure 3(c)) and metabolism (Figure 3(d)). A decreased structural coverage of the proteins in HCAP compared to the control sets was evident (Figure 3(a) *versus* Figure 3(b)–(d)).

In order to perform a quantitative analysis, the biosynthesis, degradation, and metabolism datasets were grouped together, yielding 416 proteins *versus* 231 proteins in HCAP (Table 3). Both groups had comparable percentages of sequences with at least partial 3D structural information, 53.6% for HCAP *versus* 55.4% for the combined set. However, over 2.5-fold less HCAP proteins (17.3%) had >75% PDB coverage in comparison with the proteins from the three other datasets (44.5%). A similar difference was observed when the numbers of aligned residues were analyzed: of the 141,369 residues in HCAP, only 29,825 (21%) align with homologous 3D structures, while, of the 222,987 residues in the combined group, 93,365 (42%) align with structures. The twofold smaller amount of structural information for the cancer-associated proteins further supports our disorder predictions.

Correlation of disorder predictions with the experimental structural data for HCAP

To compare disorder predictions directly with available structural data for the cancer-associated proteins, we selected 15 proteins with >45% disordered residues from the HCAP dataset (Table 4). Information detailing ordered 3D structure was found for only 13 fragments from seven of these

Table 3. Structural coverage of proteins from HCAP and three representative non-membrane categories

	HCAP		Biosynthesis + metabolism + degradation	
Coverage (% residues)	No. proteins	% Proteins	No. proteins	% Proteins
0–25	33	14.3	7	1.7
25–50	32	13.8	14	3.4
50–75	19	8.2	24	5.8
75–100	40	17.3	185	44.5
Total proteins	124/231	53.6	230/416	55.4

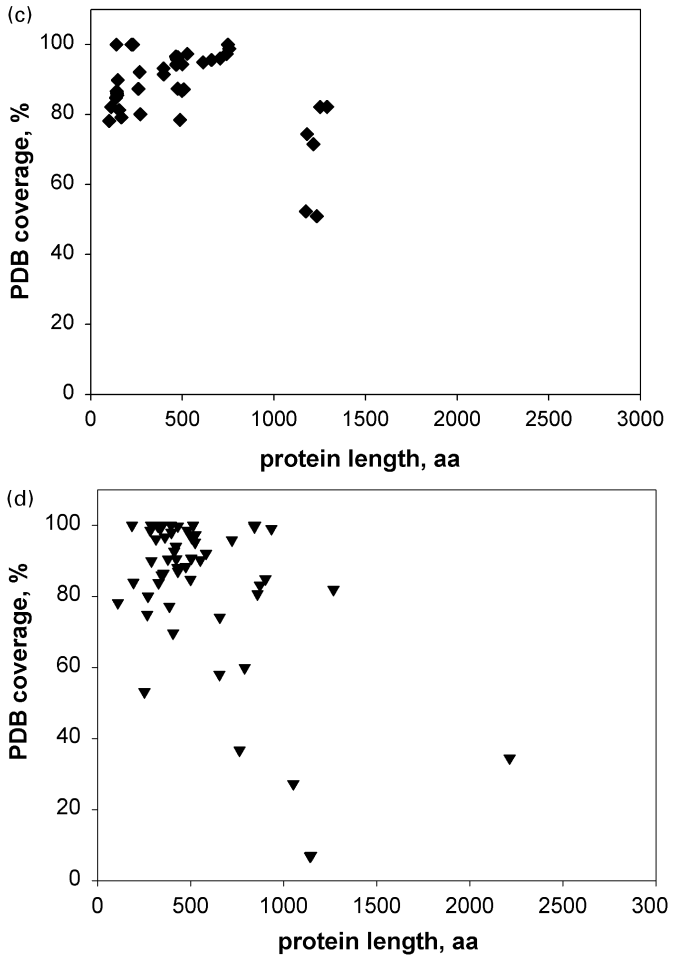

Figure 3. Structural coverage of the HCAP, biosynthesis, degradation, and metabolism datasets. Four datasets were used in a BLAST search for homologous proteins with known 3D structure as described in Materials and Methods. The *x*-axis gives the length of each protein, and the *y*-axis shows the percentage of the sequence of each protein for which the structure has been determined. The percentage coverage was calculated as the total number of residues that align with PDB structures divided by the total protein length. (a) HCAP dataset, circles; (b) biosynthesis dataset, upward triangles; (c) degradation dataset, diamonds, and (d) metabolism dataset, downward triangles.

15 proteins, comprising 882 residues of the total of 7543, or just 11.7%. Not a single structure has been solved for any of the 15 full-length proteins, despite the likelihood of numerous structure determination attempts.

Comparison of the PONDR VL-XT disorder analysis with available structural information for the seven above-noted proteins reveals that the long predictions of order correlate well with determination of 3D structure (Figure 4). The

Table 4. The 15 proteins from HCAP with >45% of residues predicted to be disordered

Protein name	SWISS-PROT accession no.	Protein length (res.)	No. dis. residues	Overall % disorder[a]	Longest DR[b]	Longest DR loc.
FRAT-1 proto-oncogene	FRT1_HUMAN	279	242	86.7	85	38–122
EWS oncogene	EWS_HUMAN	656	524	79.9	286	9–294
FUS oncogene	FUS_HUMAN	526	382	72.6	252	3–254
Cyclin-dep. kinase inhibitor p57	CDNC_HUMAN	316	226	71.5	156	109–264
AF4 proto-oncogene	AF4_HUMAN	1210	863	71.3	430	521–950
c-*jun* proto-oncogene	AP1_HUMAN	331	212	64	117	172–288
L-*myc*-1 proto-oncogene	MYCL_HUMAN	364	233	64	87	217–303
Homeobox protein Hox-11	HX11_HUMAN	330	203	61.5	111	56-166
c-*fos* proto-oncogene	FOS_HUMAN	380	232	61	107	73–179
N-*myc* proto-oncogene	MYCN_HUMAN	464	265	57.1	85	202–286
C-ski oncogene	SKI_HUMAN	728	415	57	155	421–553
Mdm2 oncoprotein	MDM2_HUMAN	491	279	56.8	81	109–189
c-*myc* proto-oncogene	MYC_HUMAN	439	247	56.3	94	203–296
Tumor protein p73	P73_HUMAN	636	349	54.9	121	367–487
Tumor suppressor p53	P53_HUMAN	393	187	47.6	66	34–99

[a] Overall percentage disorder represents the fraction of the residues predicted to be disordered.
[b] The number of residues in the longest predicted disordered region (longest DR) and the first and last residue numbers in the longest predicted DR (longest DR loc.) are given in the two right-most columns, respectively.

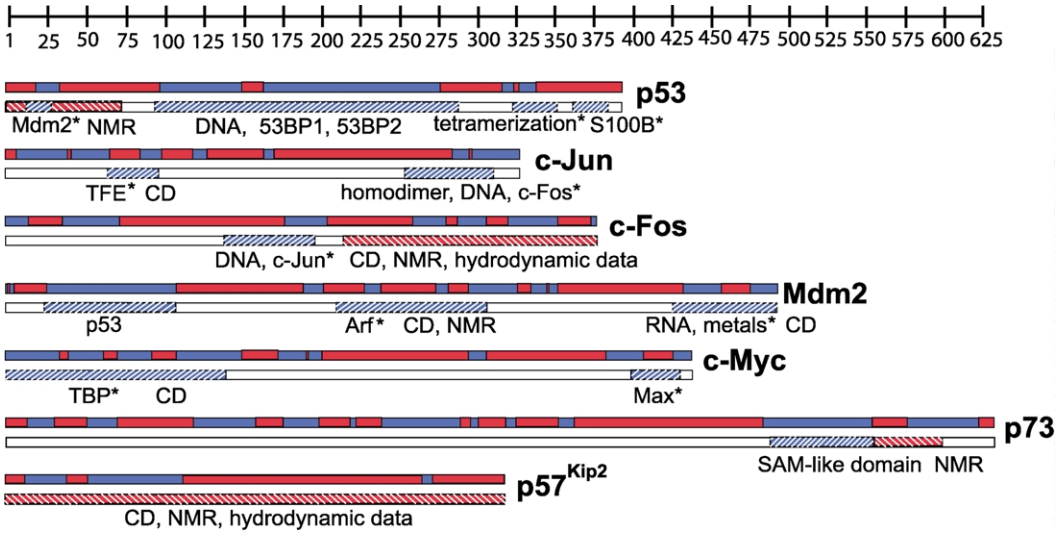

Figure 4. Comparison of PONDR VL-XT disorder predictions and experimental structural data for HCAP. The upper bar represents PONDR VL-XT predictions: red for disorder (PONDR VL-XT prediction score ≥ 0.5) and blue for order (PONDR VL-XT prediction score < 0.5). The lower bar represents experimentally verified order and disorder: red with backward slash signifies experimentally verified disorder, blue with forward slash signifies experimentally verified order, and white signifies the lack of structural information. The *x*-axis represents the residue number for each protein. The names of interacting partners and the methods used for order and disorder determination are shown under the lower bar. An asterisk ($*$) indicates that the region undergoes disorder-to-order transition upon binding to partner(s) or upon change in solvent conditions.

DNA-binding domain of p53, the fragment of Mdm2 interacting with p53, and the N-terminal part of p73 SAM-like domain all exhibited strong tendencies to be ordered by the PONDR VL-XT analysis.

Given that PONDR VL-XT was trained using segments longer than 40 consecutive amino acid residues, short regions of 3D structure show less accurate agreement with predictions. These short regions include three from p53 (the tetramerization domain, the Mdm2-binding domain, and the S100B-binding domain) and three leucine zippers (one each from c-*fos*, c-*jun* and c-*myc*) (Figure 4). PONDR VL-XT predicted a combination of ordered and disordered regions for all of these segments. Interestingly, these regions together with several others indicated by asterisks in Figure 4, all undergo disorder-to-order transitions upon oligomerization or upon binding with partners.

If the leucine zippers and the other fragments known to undergo disorder-to-order transitions are deleted from the set of ordered fragments, an overall PONDR VL-XT prediction accuracy of 90% is obtained. This value compares favorably with the 80% accuracy observed when the same predictor was applied to about 900 proteins containing about 220,000 ordered residues.[2] Thus, PONDR VL-XT predicts the regions of known 3D structure in the analyzed proteins correctly.

Just as the 3D structural information for the 15 proteins from Table 4 is very limited, the experimental data for the lengths and locations of their disordered regions is also sparse (Figure 4). NMR and CD studies indicate disorder for 580 residues, corresponding to ~8% of the total of 7543 amino acid residues in these 15 proteins. One of them, $p57^{Kip2}$, which is involved in cell-cycle arrest by inhibiting cyclin-dependent kinases, was found by CD, NMR and hydrodynamic methods to be disordered completely,[36] while another, p53, shows disordered tails by NMR of the full-length protein.[37]

Of the 580 disordered residues, 546 have been shown to be involved in coupled folding and binding. For example, the polypeptide corresponding to the c-*myc* transactivation domain showed a random conformation as determined by CD[38] until it interacted specifically with TBP, and this binding was accompanied by induction and stabilization of the secondary structure in the polypeptide. Other disordered segments such as the Arf interacting region and the RING finger domain of Mdm2 undergo similar transitions from random coil-like conformation to regular secondary structures.[39,40] The C-terminal domain of the transcription factor c-*fos*[18] is intrinsically disordered in the absence of interacting partners. For long protein fragments, such as the C terminus of c-*fos*, it is frequently unclear whether the observed disorder is intrinsic or results from the absence of stabilizing tertiary contacts.

The accuracy of PONDR VL-XT for the characterized regions of disorder was 64%, which is comparable to the 63% estimated from over 140 proteins containing more than 17,000 disordered residues.[2] The lower level of accuracy of disorder *versus* order predictions has multiple cause, as discussed elsewhere in more detail.[41] In brief, the characterization of order by X-ray diffraction and NMR is unambiguous, with atomic coordinates

being assigned for each ordered residue, whereas the characterization of disorder is usually ambiguous. For example, X-ray-characterized disorder corresponds to regions with missing electron density, but such regions can result from ordered, wobbly domains. Protein regions characterized by NMR frequently contain local segments that undergo disorder-to-order transitions upon binding with a partner. The predictions of order in such disordered regions are probably not true errors, but rather anticipate the formation of order upon binding.[42] Finally, when using CD for the global structure estimates, the signals from ordered regions can be obscured by the signals from disordered regions, and the CD-characterized disorder is expected to contain a considerable fraction of ordered residues. Overall, this ambiguity in disorder characterization is probably the most important factor leading to the lower apparent prediction accuracy for regions of disorder.

Importance of disorder for cell signaling and cancer

What is the significance of our observations? Is there any advantage to being disordered for the biological functions performed by cancer-associated and signaling proteins? Intrinsically unstructured proteins are involved frequently in numerous processes in the cell: transcriptional activation, cell-cycle regulation, membrane transport, molecular recognition and signaling.[1] The lack of folded structure might give these proteins functional advantage over globular proteins with well-defined 3D structure: the ability to bind to multiple different targets without sacrificing specificity. Moreover, intrinsic disorder might be responsible for the binding diversity of the proteins involved in the broad cascade of protein–protein interactions. Therefore, the amount of intrinsic disorder in highly connected proteins would be expected to correlate with the number of their interacting partners. Our disorder predictions on several proteins that form hubs and nodes in signaling networks (data not shown) are consistent with this suggestion.

Eukaryotic cell-signaling proteins carry numerous post-translational modifications that occur frequently in disordered regions.[2] The regulation of the c-*src* tyrosine kinase is controlled *via* tyrosine phosphorylation in the activation loop: in the inactive conformation, the loop is ordered and tyrosine is not phosphorylated, whereas upon kinase activation the loop becomes flexible and disordered facilitating the exposure of tyrosine for phosphorylation.[43] The majority of sites in p53 that are phosphorylated by casein and protein kinases are located in the putatively disordered N-terminal transactivation domain and C-terminal basic tail.[37,44] Unstructured regions of 4E-BP contain multiple phosphorylation sites that play an important role in the regulation of 4E-BP binding to eIF4E and the correlated regulation of translation by eIF4E.[45] The disordered tails of histones frequently carry a large network of post-translational modifications that are crucial for differential regulation of chromatin activity.[46] Inactivating phosphorylation of the unstructured loop region of Bcl-2 by CDK leads to the loss of its anti-apoptotic activity.[47] These examples indicate that post-translational modifications occur often in regions of intrinsic disorder. Perhaps the ability to fold onto the surface contours of modifying enzymes provides a selective advantage for localizing chemical modification sites in the disordered regions.[4]

Investigation of the evolutionary rate of the yeast protein interaction network suggests that it changes rapidly, with as many as half of all protein interactions being replaced by new ones every 300 Myr.[48] Interestingly, comparison of the evolutionary rates of several protein families indicates that the disordered protein regions evolved significantly faster than the ordered regions.[49] The use of disorder for signaling interactions might facilitate the adaptability of such rapidly evolving networks.

All cellular-signaling processes demand finely tuned regulation and fast removal of some proteins from the cell. Disordered regions likely carry the signals for proteolytic degrading machinery as an integral part of their overall regulatory function. For example, the ubiquitin-mediated proteolysis of the c-Myc protein is governed by its transcriptional activation domain,[50] shown to be unstructured without its binding partner.[38]

In conclusion, protein disorder plays an important role in many key cellular processes, and it may be involved directly in mediating interactions between highly connected proteins in signaling networks.

Implications of disorder for the discovery of anti-cancer drugs

Combining and integrating bio- and chemo-informatics promises to open new perspectives in the drug discovery process, from the identification of novel targets to the development of lead compounds with desired properties.[51] Current structure-based drug design strategies,[52,53] however, do not employ information on intrinsic disorder. Predictive algorithms such as PONDR could identify disordered regions that are very unlikely either to crystallize or to bind drug molecules by the traditional lock-and-key mechanisms. Such information would be extremely useful at the early stages of target selection. Furthermore, disorder predictors can help identify local domains within longer regions of disorder that would be amenable to structure determination, similar to the domains in p53, Mdm2, and p73 (Figure 4). Combining predictions of intrinsic disorder with other techniques provides an alternative strategy for protein structural characterization. For example, we used

PONDR in combination with limited proteolysis[54] and mass spectrometry[55] to characterize the disordered regions in two proteins, clusterin and XPA.

The development of new approaches to discover drug molecules that target intrinsically disordered protein regions should be a high priority. The important anti-cancer drug taxol, which was discovered in a random screen,[56] may act by inducing tubulin polymerization.[57] Since taxol binding is associated with protection of highly sensitive protease digestion sites, the tubulin-binding site likely involves a region of intrinsic disorder. In addition, taxol interacts with an intrinsically disordered region in Bcl-2[58,59] and thereby alters the apoptotic signaling pathway, perhaps by leading to enhanced Bcl-2 phosphorylation.[60,61] The common occurrence of intrinsic disorder in cancer-associated and signaling proteins and the ability of taxol to specifically bind to disordered protein regions suggest that disorder information should be employed in the development of new strategies for the discovery of anti-cancer drugs.

Materials and Methods

Sequences and datasets

(1) Human cancer-associated proteins (HCAP); the dataset of 231 HCAP was extracted from SWISS-PROT† using keywords "anti-oncogene OR oncogene OR proto-oncogene OR tumor" in the description field and "human" in the organism field.

(2) Signaling proteins (AfCS); the non-redundant dataset of 2329 proteins involved in cellular signaling, was created by the Alliance for Cellular Signaling‡.

(3) The eukaryotic fraction of SWISS-PROT (EU_SW); a non-redundant dataset of 53,630 protein sequences was extracted from SWISS-PROT by query "eukaryota" in the organism field.

(4) Ordered PDB_Select_25 (O_PDB_S25), 1138 entries; a dataset containing only the ordered parts of the proteins from PDB Select 25§, a non-homologous subset of the structures in PDB consisting of a single representative structure for protein families whose members have $<25\%$ sequence identity. O_PDB_S25 was constructed by removing the disordered regions (i.e. residues with backbone atoms that are not observed in X-ray crystal structures) from the PDB Select 25 protein sequences.

(5) A total of 11 additional datasets (Table 2) were extracted from SWISS-PROT using keywords "regulation", "cytoskeleton", "ribosomal", "membrane", "transport", "biosynthesis", "inhibitor", "kinase", "metabolism", "degradation", or "G-protein coupled receptor" combined with "human" in the organism field. They represent functional categories of human proteins involved in various cellular processes. These datasets overlap, i.e. the same protein can be present in several datasets. For example, trithorax-like protein HRX can be found in both HCAP and regulation datasets, because it is a proto-oncogene involved in acute leukemias, and at the same time it acts as transcriptional regulatory factor.

†http://www.expasy.ch/sprot
‡http://www.cellularsignaling.org
§http://www.cmbi.kun.nl/swift/pdbsel

Disorder predictor and the error rate

Predictions of intrinsic disorder in proteins were made using PONDR VL-XT.[14] Briefly, VL-XT was formed by merging three neural network predictors of disorder; one for N-terminal regions, a second for internal regions and a third for C-terminal regions. The merger was accomplished by performing overlapping predictions, followed by averaging the outputs. The VL-XT training set included disordered segments of 40 or more amino acid residues as characterized by X-ray and NMR for the predictor of the internal regions, and segments of five or more amino acid residues for the predictors of the two terminal regions. The false-positive error rate in the prediction of disorder for an ordered residue in O_PDB_S25 is 20% but it drops to 0.4% for ≥ 40 consecutive predictions of disorder. The false-negative error rate is 37% on a per residue basis when VL-XT is applied to 140 proteins (containing $>17{,}000$ residues) that have experimentally characterized disordered regions of at least 30 amino acid residues. This rate decreases to 11% for ordered regions of ≥ 40 residues. Because the false-negative error rate is greater than the false-positive error rate, VL-XT most likely underestimates the occurrence of long disordered regions in proteins.

Statistical analysis

An analysis of variability in the percentage of proteins with predicted disorder was performed by bootstrap resampling.[62] For each dataset from Table 1, 231 proteins were sampled randomly with replacement. For the functional protein categories from Table 2, the number of randomly sampled proteins for each dataset was equal to the number of proteins in the dataset. The fraction of proteins with disordered regions of a given length was determined for each sample. The datasets were sampled 1000 times, and these values were used to calculate the standard error of the fractions for each dataset. The 95% confidence intervals were calculated from the standard errors and are shown as error bars in Figures 1(a) and 2(b). Non-overlapping confidence intervals indicate that the fractions are significantly different.

Identification of putative structural homologues from PDB

The gapped-BLAST algorithm[63] was used to compare sequences in PDB with those in the various datasets. The filter for low sequence complexity was turned off, and the default scoring matrix (BLOSUM 62) was used. A putative structural homologue was identified if the sequence match covered at least 85% of the residues in the PDB structure with a sequence identity of at least 30%.

Acknowledgements

We thank Dr Cheryl Arrowsmith from The Ontario Cancer Institute, Dr Howard Hosick from Washington State University and, especially, D. Eric Ackerman from Pacific Northwest National Laboratories for useful discussions and critical reading of the manuscript. The anonymous reviewers were especially helpful. We thank Jason Sikes for providing expert computer programming and technical support. This study was supported by NIH grant 1R01 LM 06916, and NSF grants CSE-IIS-9711532 and CSE-IIS-0196237.

References

1. Wright, P. E. & Dyson, H. J. (1999). Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm. *J. Mol. Biol.* **293**, 321–331.
2. Dunker, A. K., Lawson, J. D., Brown, C. J., Williams, R. M., Romero, P., Oh, J. S. *et al.* (2001). Intrinsically disordered protein. *J. Mol. Graph. Model.* **19**, 26–59.
3. Uversky, V. N. (2002). What does it mean to be natively unfolded? *Eur. J. Biochem.* **269**, 2–12.
4. Dunker, A. K., Brown, C. J., Lawson, J. D., Iakoucheva, L. M. & Obradović, Z. (2002). Intrinsic disorder and protein function. *Biochemistry,* **41**, 6573–6582.
5. Spolar, R. S. & Record, M. T., II (1994). Coupling of local folding to site-specific binding of proteins to DNA. *Science*, **263**, 777–784.
6. Demchenko, A. P. (2001). Recognition between flexible protein molecules: induced and assisted folding. *J. Mol. Recogn.* **14**, 42–61.
7. Dyson, H. J. & Wright, P. E. (2002). Coupling of folding and binding for unstructured proteins. *Curr. Opin. Struct. Biol.* **12**, 54–60.
8. Schulz, G. E. (1979). Nucleotide binding proteins. In *Molecular Mechanism of Biological Recognition* (Balaban, M., ed.), pp. 79–94, Elsevier/North-Holland, New York.
9. Kriwacki, R. W., Hengst, L., Tennant, L., Reed, S. I. & Wright, P. E. (1996). Structural studies of $p21^{Waf1/Cip1/Sdi1}$ in the free and Cdk2-bound state: conformational disorder mediates binding diversity. *Proc. Natl Acad. Sci. USA*, **93**, 11504–11509.
10. Dunker, A. K. & Obradović, Z. (2001). The protein trinity-linking function and disorder. *Nature Biotechnol.* **19**, 805–806.
11. Chervitz, S. A., Aravind, L., Sherlock, G., Ball, C. A., Koonin, E. V., Dwight, S. S. *et al.* (1998). Comparison of the complete protein sets of worm and yeast: orthology and divergence. *Science*, **282**, 2022–2028.
12. Dunker, A. K., Obradović, Z., Romero, P., Garner, E. C. & Brown, C. J. (2000). Intrinsic protein disorder in complete genomes. *Genome Inform.* **11**, 161–171.
13. Sigler, P. B. (1988). Transcriptional activation. Acid blobs and negative noodles. *Nature*, **333**, 210–212.
14. Romero, P., Obradović, Z., Li, X., Garner, E. C., Brown, C. J. & Dunker, A. K. (2001). Sequence complexity of disordered protein. *Proteins: Struct. Funct. Genet.* **42**, 38–48.
15. Hartwell, L. H. & Kastan, M. B. (1994). Cell cycle control and cancer. *Science*, **266**, 1821–1828.
16. Fearon, E. R. & Vogelstein, B. (1990). A genetic model for colorectal tumorigenesis. *Cell*, **61**, 759–767.
17. DiGiammarino, E. L., Filippov, I., Weber, J. D., Bothner, B. & Kriwacki, R. W. (2001). Solution structure of the p53 regulatory domain of the p19Arf tumor suppressor protein. *Biochemistry,* **40**, 2379–2386.
18. Campbell, K. M., Terrell, A. R., Laybourn, P. J. & Lumb, K. J. (2000). Intrinsic structural disorder of the C-terminal activation domain from the bZIP transcription factor Fos. *Biochemistry,* **39**, 2708–2713.
19. Metz, R., Bannister, A. J., Sutherland, J. A., Hagemeier, C., O'Rourke, E. C., Cook, A. *et al.* (1994). c-Fos-induced activation of a TATA-box-containing promoter involves direct contact with TATA-box-binding protein. *Mol. Cell Biol.* **14**, 6021–6029.
20. Bannister, A. J. & Kouzarides, T. (1995). CBP-induced stimulation of c-Fos activity is abrogated by E1A. *EMBO J.* **14**, 4758–4762.
21. Zhang, Y., Feng, X. H. & Derynck, R. (1998). Smad3 and Smad4 cooperate with c-Jun/c-Fos to mediate TGF-beta-induced transcription. *Nature,* **394**, 909–913.
22. Demarest, S. J., Martinez-Yamout, M., Chung, J., Chen, H., Xu, W., Dyson, H. J. *et al.* (2002). Mutual synergistic folding in recruitment of CBP/p300 by p160 nuclear receptor coactivators. *Nature*, **415**, 549–553.
23. Lupas, A., Van Dyke, M. & Stock, J. (1991). Predicting coiled coils from protein sequences. *Science*, **252**, 1162–1164.
24. Namba, K. (2001). Roles of partly unfolded conformations in macromolecular self-assembly. *Genes Cells*, **6**, 1–12.
25. Gundersen, G. G. & Cook, T. A. (1999). Microtubules and signal transduction. *Curr. Opin. Cell Biol.* **11**, 81–94.
26. Sastry, S. K. & Burridge, K. (2000). Focal adhesions: a nexus for intracellular signaling and cytoskeletal dynamics. *Expt. Cell. Res.* **261**, 25–36.
27. Bauch, A., Alt, F. W., Crabtree, G. R. & Snapper, S. B. (2000). The cytoskeleton in lymphocyte signaling. *Advan. Immunol.* **75**, 89–114.
28. Fox, J. E. (2001). Cytoskeletal proteins and platelet signaling. *Thromb. Haemost.* **86**, 198–213.
29. Behrens, J. (1999). Cadherins and catenins: role in signal transduction and tumor progression. *Cancer Metastasis Rev.* **18**, 15–30.
30. Ban, N., Nissen, P., Hansen, J., Capel, M., Moore, P. B. & Steitz, T. A. (1999). Placement of protein and RNA structures into a 5 Å-resolution map of the 50 S ribosomal subunit. *Nature*, **400**, 841–847.
31. Ban, N., Nissen, P., Hansen, J., Moore, P. B. & Steitz, T. A. (2000). The complete atomic structure of the large ribosomal subunit at 2.4 Å resolution. *Science*, **289**, 905–920.
32. Williams, R. J. P. (1979). The conformational properties of proteins in solution. *Biol. Rev. Camb. Phil. Soc.* **54**, 389–437.
33. Uversky, V., Gillespie, J. & Fink, A. (2000). Why are "natively unfolded" proteins unstructured under physiologic conditions? *Proteins: Struct. Funct. Genet.* **41**, 415–427.
34. Romero, P., Obradović, Z., Kissinger, C. R., Villafranca, J. E. & Dunker, A. K. (1997). Identifying disordered regions in proteins from amino acid sequences. *IEEE Int. Conf. Neural Netw.* **1**, 90–95.
35. Romero, P., Obradović, Z., Kissinger, C. R., Villafranca, J. E., Guilliot, S., Garner, E. *et al.* (1998).

Thousands of proteins likely to have long disordered regions. *Pac. Symp. Biocomput.* **3**, 437–448.
36. Adkins, J. N. & Lumb, K. J. (2002). Intrinsic structural disorder and sequence features of the cell cycle inhibitor p57(Kip2). *Proteins: Struct. Funct. Genet.* **46**, 1–7.
37. Ayed, A., Mulder, F. A., Yi, G. S., Lu, Y., Kay, L. E. & Arrowsmith, C. H. (2001). Latent and active p53 are identical in conformation. *Nature Struct. Biol.* **8**, 756–760.
38. McEwan, I. J., Dahlman-Wright, K., Ford, J. & Wright, A. P. (1996). Functional interaction of the c-Myc transactivation domain with the TATA binding protein: evidence for an induced fit model of transactivation domain folding. *Biochemistry,* **35**, 9584–9593.
39. Bothner, B., Lewis, W. S., DiGiammarino, E. L., Weber, J. D., Bothner, S. J. & Kriwacki, R. W. (2001). Defining the molecular basis of Arf and Hdm2 interactions. *J. Mol. Biol.* **314**, 263–277.
40. Lai, Z., Freedman, D. A., Levine, A. J. & McLendon, G. L. (1998). Metal and RNA binding properties of the hdm2 RING finger domain. *Biochemistry,* **37**, 17005–17015.
41. Dunker, A. K., Brown, C. J. & Obradović, Z. (2002). Identification and functions of usefully disordered proteins. *Advan. Protein Chem.* **62**, 25–49.
42. Garner, E., Romero, P., Dunker, A. K., Brown, C. & Obradović, Z. (1999). Predicting binding regions within disordered proteins. *Genome Inform.* **10**, 41–50.
43. Xu, W., Doshi, A., Lei, M., Eck, M. J. & Harrison, S. C. (1999). Crystal structures of c-Src reveal features of its autoinhibitory mechanism. *Mol. Cell*, **3**, 629–638.
44. Ko, L. J. & Prives, C. (1996). p53: puzzle and paradigm. *Genes Dev.* **10**, 1054–1072.
45. Gingras, A. C., Gygi, S. P., Raught, B., Polakiewicz, R. D., Abraham, R. T., Hoekstra, M. F. *et al.* (1999). Regulation of 4E-BP1 phosphorylation: a novel two-step mechanism. *Genes Dev.* **13**, 1422–1437.
46. Nakayama, J., Rice, J. C., Strahl, B. D., Allis, C. D. & Grewal, S. I. (2001). Role of histone H3 lysine 9 methylation in epigenetic control of heterochromatin assembly. *Science*, **292**, 110–113.
47. Ojala, P. M., Yamamoto, K., Castanos-Velez, E., Biberfeld, P., Korsmeyer, S. J. & Makela, T. P. (2000). The apoptotic v-cyclin–CDK6 complex phosphorylates and inactivates Bcl-2. *Nature Cell. Biol.* **2**, 819–825.
48. Wagner, A. (2001). The yeast protein interaction network evolves rapidly and contains few redundant duplicate genes. *Mol. Biol. Evol.* **18**, 1283–1292.
49. Brown, C. J., Takayama, S., Campen, A. M., Vise, P., Marshall, T., Oldfield, C. J. *et al.* (2002). Evolutionary rate heterogeneity in proteins with long disordered regions. *J. Mol. Evol.* **55**, 104–110.
50. Salghetti, S. E., Kim, S. Y. & Tansey, W. P. (1999). Destruction of Myc by ubiquitin-mediated proteolysis: cancer-associated and transforming mutations stabilize Myc. *EMBO J.* **18**, 717–726.
51. Bajorath, J. (2001). Rational drug discovery revisited: interfacing experimental programs with bio- and chemo-informatics. *Drug Discov. Today*, **6**, 989–995.
52. Blundell, T. L. (1996). Structure-based drug design. *Nature*, **384**, 23–26.
53. Traxler, P., Bold, G., Buchdunger, E., Caravatti, G., Furet, P., Manley, P. *et al.* (2001). Tyrosine kinase inhibitors: from rational design to clinical trials. *Med. Res. Rev.* **21**, 499–512.
54. Bailey, R. W., Dunker, A. K., Brown, C. J., Garner, E. C. & Griswold, M. D. (2001). Clusterin: a binding protein with a molten globule-like region. *Biochemistry,* **40**, 11828–11840.
55. Iakoucheva, L. M., Kimzey, A. L., Masselon, C. D., Bruce, J. E., Garner, E. C., Brown, C. J. *et al.* (2001). Identification of intrinsic order and disorder in the DNA repair protein XPA. *Protein Sci.* **10**, 560–571.
56. Kingston, D. G. (1994). Taxol: the chemistry and structure–activity relationships of a novel anticancer agent. *Trends Biotechnol.* **12**, 222–227.
57. de Pereda, J. M. & Andreu, J. M. (1996). Mapping surface sequences of the tubulin dimer and taxol-induced microtubules with limited proteolysis. *Biochemistry,* **35**, 14184–14202.
58. Fang, G., Chang, B. S., Kim, C. N., Perkins, C., Thompson, C. B. & Bhalla, K. N. (1998). "Loop" domain is necessary for taxol-induced mobility shift and phosphorylation of Bcl-2 as well as for inhibiting taxol-induced cytosolic accumulation of cytochrome *c* and apoptosis. *Cancer Res.* **58**, 3202–3208.
59. Rodi, D. J., Janes, R. W., Sanganee, H. J., Holton, R. A., Wallace, B. A. & Makowski, L. (1999). Screening of a library of phage-displayed peptides identifies human bcl-2 as a taxol-binding protein. *J. Mol. Biol.* **285**, 197–203.
60. Haldar, S., Jena, N. & Croce, C. M. (1995). Inactivation of Bcl-2 by phosphorylation. *Proc. Natl Acad. Sci. USA*, **92**, 4507–4511.
61. Blagosklonny, M. V. (2001). Unwinding the loop of Bcl-2 phosphorylation. *Leukemia*, **15**, 869–874.
62. Efron, B. & Tibshirani, R. (1993). *An Introduction to the Bootstrap*, Chapman & Hall, New York.
63. Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W. *et al.* (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucl. Acids Res.* **25**, 3389–3402.

Edited by P. Wright

(Received 11 April 2002; received in revised form 25 July 2002; accepted 23 August 2002)