Intrinsic Disorder in Cell-signaling and Cancer-associated Proteins

Lilia M. Iakoucheva1, Celeste J. Brown1, J. David Lawson1 Zoran Obradović2 and A. Keith Dunker1*

1Department of Biochemistry and Biophysics, School of Molecular Biosciences Washington State University Pullman, WA 99164-4660 USA
2Center for Information Science and Technology, Temple University, Philadelphia, PA 19122-6094, USA

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 $79(\pm 5)\%$, $66(\pm 6)\%$, $47(\pm 4)\%$ and $13(\pm 4)\%$ 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.

*Corresponding author

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

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

Abbreviations used: CDK, cyclin-dependent kinase; eIF4E, translation initiation factor (eIF) 4E; 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
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. Second, intrinsic disorder promotes binding diversity by enabling proteins to interact with numerous partners. 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.

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. This re-use of domains suggests that multicellular organisms have developed elaborate signal transduction and regulatory control by employing novel proteins.

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) to systemati-cally 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.

The analysis of PONDR VL-XT predictions demonstrates that predicted disorder followed the ranking \( \text{HCAP} > \text{AFCS} > \text{EU_SW} > \text{O_PDB_S25} \) (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_PDB_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, 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. 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 oncogenes further supports 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. The C-terminal activation domain of c-fos in its biologically active form is structurally disordered. This domain interacts directly with multiple transcription factors: TBP, TFIIH, CBP and Smad3, and activates transcription in different cellular processes. Disorder would be a significant factor contributing to the conformational freedom of this domain and...
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. 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
Moreover, for regions of length of proteins from this category (Figure 2(a)).

The high ratio of charged to hydrophobic amino acid residues has been suggested as the likely cause of the observed disorder in these proteins. 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

---

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

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

Proteins shorter than 30 amino acid residues (res.) were eliminated from the PONDRL-VL-XT predictions.
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 $\geq$ the value on the x-axis.
(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.

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

<table>
<thead>
<tr>
<th>Coverage (%)</th>
<th>HCAP</th>
<th>Biosynthesis + metabolism + degradation</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. proteins</td>
<td>% Proteins</td>
<td>No. proteins</td>
</tr>
<tr>
<td>0–25</td>
<td>33</td>
<td>14.3</td>
</tr>
<tr>
<td>25–50</td>
<td>32</td>
<td>13.8</td>
</tr>
<tr>
<td>50–75</td>
<td>19</td>
<td>8.2</td>
</tr>
<tr>
<td>75–100</td>
<td>40</td>
<td>17.3</td>
</tr>
<tr>
<td>Total proteins</td>
<td>124/231</td>
<td>53.6</td>
</tr>
</tbody>
</table>

**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
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

<table>
<thead>
<tr>
<th>Protein name</th>
<th>SWISS-PROT accession no.</th>
<th>Protein length (res.)</th>
<th>No. dis. residues</th>
<th>Overall % disorder</th>
<th>Longest DR</th>
<th>Longest DR loc.</th>
</tr>
</thead>
<tbody>
<tr>
<td>FRAT-1 proto-oncogene</td>
<td>FRT1_HUMAN</td>
<td>279</td>
<td>242</td>
<td>86.7</td>
<td>85</td>
<td>38–122</td>
</tr>
<tr>
<td>EWS oncogene</td>
<td>EWS_HUMAN</td>
<td>656</td>
<td>524</td>
<td>79.9</td>
<td>286</td>
<td>9–284</td>
</tr>
<tr>
<td>FUS oncogene</td>
<td>FUS_HUMAN</td>
<td>526</td>
<td>382</td>
<td>72.6</td>
<td>252</td>
<td>3–254</td>
</tr>
<tr>
<td>Cyclin-dep. kin. inhib. p57</td>
<td>CDNC_HUMAN</td>
<td>316</td>
<td>226</td>
<td>71.5</td>
<td>156</td>
<td>109–264</td>
</tr>
<tr>
<td>AF4 proto-oncogene</td>
<td>AF4_HUMAN</td>
<td>1210</td>
<td>863</td>
<td>71.3</td>
<td>430</td>
<td>521–950</td>
</tr>
<tr>
<td>c-jun proto-oncogene</td>
<td>AP1_HUMAN</td>
<td>331</td>
<td>212</td>
<td>64</td>
<td>117</td>
<td>172–288</td>
</tr>
<tr>
<td>L-myc-l proto-oncogene</td>
<td>MYCL_HUMAN</td>
<td>364</td>
<td>233</td>
<td>64</td>
<td>87</td>
<td>217–303</td>
</tr>
<tr>
<td>Homeobox protein Hox-11</td>
<td>HX11_HUMAN</td>
<td>330</td>
<td>203</td>
<td>61.5</td>
<td>111</td>
<td>56–166</td>
</tr>
<tr>
<td>c-fos proto-oncogene</td>
<td>FOS_HUMAN</td>
<td>380</td>
<td>232</td>
<td>61</td>
<td>107</td>
<td>73–179</td>
</tr>
<tr>
<td>N-myc proto-oncogene</td>
<td>MYCN_HUMAN</td>
<td>464</td>
<td>265</td>
<td>57.1</td>
<td>85</td>
<td>202–286</td>
</tr>
<tr>
<td>C-ski oncogene</td>
<td>SKL_HUMAN</td>
<td>728</td>
<td>415</td>
<td>57</td>
<td>155</td>
<td>421–553</td>
</tr>
<tr>
<td>Mdm2 oncoprotein</td>
<td>MDM2_HUMAN</td>
<td>491</td>
<td>279</td>
<td>56.8</td>
<td>81</td>
<td>109–189</td>
</tr>
<tr>
<td>c-myc proto-oncogene</td>
<td>MYC_HUMAN</td>
<td>439</td>
<td>247</td>
<td>56.3</td>
<td>94</td>
<td>203–296</td>
</tr>
<tr>
<td>Tumor protein p73</td>
<td>P73_HUMAN</td>
<td>636</td>
<td>349</td>
<td>54.9</td>
<td>121</td>
<td>367–487</td>
</tr>
<tr>
<td>Tumor suppressor p53</td>
<td>P53_HUMAN</td>
<td>393</td>
<td>187</td>
<td>47.6</td>
<td>66</td>
<td>34–99</td>
</tr>
</tbody>
</table>

* Overall percentage disorder represents the fraction of the residues predicted to be disordered.

* 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.
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. Thus, PONDR VL-XT predicts the regions of known 3D structure in the analyzed proteins correctly.

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. 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, p57Kip2, which is involved in cell-cycle arrest by inhibiting cyclin-dependent kinases, was found by CD, NMR and hydrodynamic methods to be disordered completely, while another, p53, shows disordered tails by NMR of the full-length protein.

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 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.

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. The lower level of accuracy of disorder versus order predictions has multiple causes, as discussed elsewhere in more detail. In brief, the characterization of order by X-ray diffraction and NMR is unambiguous, with atomic coordinates

![Figure 4](image-url)
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. 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. The regulation of the c-Myc 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. 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. 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. The disordered tails of histones frequently carry a large network of post-translational modifications that are crucial for differential regulation of chromatin activity. Inactivating phosphorylation of the unstructured loop region of Bcl-2 by CDK leads to the loss of its anti-apoptotic activity. 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.

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. Interestingly, comparison of the evolutionary rates of several protein families indicates that the disordered protein regions evolved significantly faster than the ordered regions. 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, shown to be unstructured without its binding partner.

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 chemoinformatics 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. Current structure-based drug design strategies, 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\textsuperscript{54} and mass spectrometry\textsuperscript{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,\textsuperscript{56} may act by inducing tubulin polymerization.\textsuperscript{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\textsuperscript{58,59} and thereby alters the apoptotic signaling pathway, perhaps by leading to enhanced Bcl-2 phosphorylation.\textsuperscript{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\textsuperscript{†} 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.\textsuperscript‡

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\textsuperscript§, 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 a transcriptional regulatory factor.

**Disorder predictor and the error rate**

Predictions of intrinsic disorder in proteins were made using PONDR VL-XT.\textsuperscript{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.\textsuperscript{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-\textsc{blast} algorithm\textsuperscript{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 (\textsc{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\%.

\textsuperscript{†}http://www.expasy.ch/sprot

\textsuperscript{‡}http://www.cellularsignaling.org

\textsuperscript{§}http://www.cmbi.kun.nl/swift/pdbsel
Acknowledgements

We thank Dr. Cheryl Arrowsmith from The Ontario Cancer Institute, Dr. Howard Hosick from Washington State University and, especially, Dr. 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


Edited by P. Wright

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