# Intrinsic Disorder in Transcription Factors[†]

Jiangang Liu,[‡,||] Narayanan B. Perumal,[‡] Christopher J. Oldfield,[‡,§] Eric W. Su,[||] Vladimir N. Uversky,*[,§,⊥,#] and A. Keith Dunker[‡,§]

*School of Informatics, Indiana University-Purdue University Indianapolis, 535 West Michigan Street, Indianapolis, Indiana 46202, Department of Biochemistry and Molecular Biology, and the Center for Computational Biology and Bioinformatics, Indiana University School of Medicine, 714 North Senate Avenue, Suite 250, Indianapolis, Indiana 46202, Bioinformatics Group, Lilly Research Laboratories, Eli Lilly and Company, DC GL54 Greenfield, Indiana 46140, Institute for Biological Instrumentation, Russian Academy of Sciences, 142292 Pushchino, Moscow Region, Russia, and Molecular Kinetics, Inc., 6201 La Pas Trail, Suite 160, Indianapolis, Indiana 46268*

*Received February 8, 2006; Revised Manuscript Received March 8, 2006*

ABSTRACT: Intrinsic disorder (ID) is highly abundant in eukaryotes, which reflect the greater need for disorder-associated signaling and transcriptional regulation in nucleated cells. Although several well-characterized examples of intrinsically disordered proteins in transcriptional regulation have been reported, no systematic analysis has been reported so far. To test for the general prevalence of intrinsic disorder in transcriptional regulation, we used the predictor of natural disorder regions (PONDR) to analyze the abundance of intrinsic disorder in three transcription factor datasets and two control sets. This analysis revealed that from 94.13 to 82.63% of transcription factors possess extended regions of intrinsic disorder, relative to 54.51 and 18.64% of the proteins in two control datasets, which indicates the significant prevalence of intrinsic disorder in transcription factors. This propensity of transcription factors to intrinsic disorder was confirmed by cumulative distribution function analysis and charge-hydropathy plots. The amino acid composition analysis showed that all three transcription factor datasets were substantially depleted in order-promoting residues and significantly enriched in disorder-promoting residues. Our analysis of the distribution of disorder within the transcription factor datasets revealed that (a) the AT-hooks and basic regions of transcription factor DNA-binding domains are highly disordered; (b) the degree of disorder in transcription factor activation regions is much higher than that in DNA-binding domains; (c) the degree of disorder is significantly higher in eukaryotic transcription factors than in prokaryotic transcription factors; and (d) the level of α-MoRF (molecular recognition feature) prediction is much higher in transcription factors. Overall, our data reflected the fact that eukaryotes with well-developed gene transcription machinery require transcription factor flexibility to be more efficient.

Recent evidence suggests that eukaryotic genomes are highly enriched in intrinsic disorder (ID)[1] proteins relative to bacteria and archaea (*1*−*3*), which may reflect the greater need for signaling and regulation in nucleated cells (*4*−*6*). ID regions promote molecular recognition primarily through four features (*7*). (1) ID proteins or ID regions are characterized by the unique combination of high specificity and low affinity in their interactions with functional partners, which is very important for transient protein−protein and protein−nucleic acid interactions, such as those that frequently occur during the signal transduction, recognition, and regulation events. (2) Intrinsic plasticity enables a single ID protein or region to recognize and bind many biological targets with high specificity (*4*, *6*, *8*, *9*). (3) The ID proteins or regions have the propensity to form large interaction surfaces allowing them to wrap-up or surround their binding partners (*4*, *10*, *11*). (4) The rapid turnover and reduced lifetime of ID proteins in the cell might represent an important regulatory mechanism (*8*, *12*).

The fact that several experimentally well-characterized proteins, such as p53, GCN4, CBP, HMGA, and BRCA1, interact with many partners via ID regions strongly support this concept (*6*, *7*, *9*, *13*−*16*). Furthermore, the number of ID proteins known to be involved in cell signaling and regulation is growing rapidly. To better understand the prevalence of ID, several attempts have been made to apply different ID predictors in a genome-wide scale or to large protein databases. For example, Iakoucheva and her colleagues analyzed several datasets extracted from Swiss-Prot using a number of order−disorder predictors and discovered that ID is prevalent in cell-signaling and cancer-associated proteins in comparison with other functional groups (*5*). The

[†] This work was supported in part by the Indiana Genomics Initiative (INGEN) and NIH Grant R01 LM007688-01A1 (to A.K.D.) and the Programs of the Russian Academy of Sciences for the Molecular and Cellular Biology and Fundamental Science for Medicine (to V.N.U.).

* To whom correspondence should be addressed. Tel: 317-278-9650. Fax: 317-278-9217. E-mail: vuversky@iupui.edu.

[‡] Indiana University-Purdue University Indianapolis.

[§] Indiana University School of Medicine.

[||] Eli Lilly and Company.

[⊥] Russian Academy of Sciences.

[#] Molecular Kinetics, Inc.

[1] Abbreviations: TF, transcriptional factor; ID, intrinsic disorder; α-MoRE, α-helix-forming molecular recognition element; α-MoRF, α-helix-forming molecular recognition feature; PONDR, predictor of natural disordered regions (PONDR is a registered trademark of Molecular Kinetics, Inc.); PDB, protein data bank.

10.1021/bi0602718 CCC: $33.50
Published on Web 05/13/2006

results obtained by a different group for the *Saccharomyces* genome suggest that the proteins containing disorder are over-represented in the cell nucleus and are likely to be involved in the regulation of transcription and cell signaling (*2*). The results also indicate that ID is often associated with such molecular functions as kinase activity and nucleic acid binding.

Transcription factors (TFs) function through the recognition of specific DNA sequences and recruitment and assembly of the transcription machinery. In this sense, both protein–DNA and protein–protein recognition are central processes in TF function. It has been reported that protein–protein and protein–DNA interactions are often accompanied by a local folding in a protein molecule (*17*). One of the important biological implications of this coupled binding and folding scenario is that protein backbone mobility may play an important role in the early stages of a binding event (*18*), where the specific signal from the complex of protein with its binding partner emerges only after appropriate conformational changes take place (*19*). In support of this idea, it has been shown that the high degree of backbone mobility of the *lac* repressor facilitates its association with nonspecific DNA, but the binding to specific DNA is accompanied by a large decrease in backbone mobility (*20*). Available evidence, therefore, points to a central role of ID in the function of TFs.

Recently, a specific structural element called MoRE (molecular recognition element or MoRF, molecular recognition feature) has been proposed to function in the recognition of protein or nucleic acid partners (*21*). This element consists of a short region (on the order of 20 residues) that undergoes a disorder-to-order transition that is stabilized by binding to its partner (*21*, *22*). It has been reported that the frequency of α-MoRFs in various types of proteins was highest in those associated with signaling and lowest in the metabolic enzymes (*21*).

Although several well-characterized examples of the individual ID proteins involved in transcriptional regulation have been described in the literature (*6*, *9*), no systematic analysis of the abundance of ID in gene regulation proteins has been reported so far. Given the established role of disorder in a small number of characterized TFs, it is reasonable to assume that disorder may be an important and prevalent feature in all TFs. To check this hypothesis, we predicted disorder for three TF datasets and two control sets using two predictors of natural disordered regions (PONDR), VL-XT (*23*) and VSL1 (*24*), and cumulative distribution function (CDF) (*3*) and charge-hydropathy plot (CH-plot) analyses (*25*). We also surveyed the local and overall amino acid composition biases observed in TFs. Furthermore, a detailed computational analysis of the unstructured regions associated with different TF domains and subdomains has been performed. The differences between the DNA-binding and trans-activation domains and between the TF regions that bind DNA major and minor grooves have been analyzed. The specific structural element, α-MoRE (*21*), responsible for protein–protein or protein–nucleic acid interactions in TFs have been examined. The predicted properties of TF domains and subdomains have been compared with experimental results from other studies.

## MATERIALS AND METHODS

### *Datasets*

Four different data sets have been created and used for the studies as described below.

*PDBs25*. This dataset contains 1771 chains with 297 372 residues and is a nonhomologous subset of the structures in the PDB consisting of a single representative structure for protein families whose members have <25% sequence identity (*26*).

*TF_SP_NR25, TF_SP&TRE_NR25, TF_NR25, and RANDOMAC_NR25*. To construct the nonredundant, representative datasets for TFs from Swiss-Prot, 2683 protein sequences were downloaded only from Swiss-Prot (for the TF_SP_NR25 dataset), and total of 7195 entries were retrieved from both Swiss-Prot and TrEMBL (Swiss-Prot Release 46.2 of 01-Mar-2005, 172 233 entries; TrEMBL Release 29.2 of 01-Mar-2005, 1 631 173 entries) by using TF as a key word in a full-text search (for the TF_SP&TRE_NR25). The third dataset for TF_NR25 contains 1186 protein sequences and was retrieved from the TRANSFAC database (TRANSCRIPTION FACTOR TABLE, Release 3.2 pf 26-06-1997) (http://www.generegulation.com/pub/databases.html#transfac) on the basis of the availability of both sequence and domain features. RandomAC_NR25 was built with randomized the NCBI (GenBank) accession number. Four thousand six digit GenBank accession numbers were generated randomly and then used to retrieve 2387 sequences. The redundancy among sequences in these four datasets was first removed by using the CD-HIT (cluster database at high identity with tolerance) program from http://bioinformatics.org/project/?group_id=350 (*27*, *28*) to reduce the homology to 80%, and then to 40% sequence identity according to the recommended procedures. To have no two sequences in the resulting dataset with more than 25% sequence identity, we aligned these sequences against one another using a global pairwise sequence alignment program called stretcher (http://emboss.sourceforge.net/apps/stretcher.html). Any two sequences with an identity >25% will be stripped off from the dataset.

### *Disorder Predictions*

Prediction of ID in TFs was performed using PONDR VL-XT (*23*, *29*–*31*)(http://www.pondr.com), CDF (*1*), and charge-hydropathy plots (*25*).

*PONDRVL-XT*. PONDR (predictor of natural disordered regions) is a set of neural network predictors of disordered regions on the basis of local amino acid composition, flexibility, hydropathy, coordination number, and other factors. These predictors classify each residue within a sequence as either ordered or disordered. VL-XT integrates three feed forward neural networks: the variously characterized long, version 1 (VL1) predictor from Romero et al. 2000, which predicts nonterminal residues, and the X-ray characterized *N*- and *C*-terminal predictors (XT) from Li et al. 1999 (*29*), which predicts terminal residues. The output of the XT predictor provides predictions up to 14 amino acids from their respective ends. A simple average is taken for the overlapping predictions, and a sliding window of nine amino acids is used to smooth the prediction values along the length of the sequence. Unsmoothed prediction values from the XT predictors are used for the first and last four sequence positions.

*PONDR VSL1 Predictions*. The recently developed Various Short-Long, version 1 (VSL1) algorithm is an ensemble of logistic regression models that predict per-residue order–disorder (*24*, *32*). Two models predict either long or short disordered regions, greater or less than 30 residues, on the basis of features similar to those used by VL-XT. The algorithm calculates a weighted average of these predictions, where the weights are determined by a meta predictor that approximates the likelihood of a long disordered region within its 61-residue window. Predictor inputs include PSI-blast (*33*), profiles and PHD (*34*), and PSI-pred (*35*) secondary structure predictions. To reduce the time required to calculate these predictions, we omitted secondary structure prediction inputs, which decrease the overall accuracy of predictions by only ~1% (Peng, K., personal communication). Predictions were run on the AVIDD-I cluster at Indiana University-Perdue University at Indianapolis.

*Cumulative Distribution Functions (CDFs)*. 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; therefore, disordered and wholly disordered proteins tend to lie on either side of this boundary. Alternatively, the prediction can be displayed as a histogram. From each histogram, a cumulative distribution function (CDF) (*36*), can be calculated by determining the fraction of the distribution that lies below a given value (*1*, *3*). However, this method summarizes these per-residue predictions by plotting PONDR scores against their cumulative frequency, which allows ordered and disordered proteins to be distinguished on the basis of the distribution of prediction scores.

*Charge-Hydropathy Plots*. Another established method of order–disorder classification is charge-hydropathy plots (*25*). Ordered and disordered proteins plotted in charge-hydropathy space can be separated to a significant degree by a linear boundary. The normalized hydropathy of each amino acid sequence was calculated by the Kyte–Doolittle approximation within a window size of 5 amino acids. The mean hydropathy is defined as the sum of the normalized hydropathies of all residues divided by the number of residues in the polypeptide. The mean net charge is defined as the sum of basic residues minus the sum of acidic residues, divided by the total number of residues. The absolute value of the return is the formal euclidian distance of a protein in charge/hydropathy space from a previously calculated order/disorder boundary. The sign of the return is positive, if the protein is disordered (above the boundary) or negative, if the protein is ordered (below the boundary).

*α-MoRF Predictions*. The development of a preliminary α-MoRE (α-MoRF) predictor has been described previously (*21*). This predictor is based on observations that predictions of order in otherwise highly disordered proteins corresponds to protein regions that mediate interaction with other proteins or DNA (*21*, *37*). This predictor focuses on short binding regions within long regions of disorder that are likely to form helical structure upon binding. It uses a stacked architecture, where VL-XT is used to identify short predictions of order within long predictions of disorder, and then, a second level predictor determines whether the order prediction is likely to be a binding site based on attributes of both the predicted ordered region and the predicted surrounding disordered region. Recent results indicate that the α-MoRF predictor has poor sensitivity, that is, it misses many α-MoRF regions (Yugong C., unpublished data), which is not surprising given the small set of α-MoRF regions used in its development.

### *Extraction of TF Domain Information*

All of the domain information was extracted from the section Feature Table Data in each entry in the Swiss-Prot format. The FT (feature table) lines provide a precise but simple means for the annotation of the sequence data. The table describes regions or sites of interest in the sequence. In general, the feature table lists posttranslational modifications, binding sites, enzyme active sites, local secondary structure, or other characteristics reported in the cited references. The FT lines have a fixed format. The column numbers are allocated to each of the data items within each FT line. The residue positions for each feature were used to retrieve the corresponding domain sequence and PONDR VL-XT prediction score.

### *Amino Acid Composition Plots*

To compare the compositions of the three different TF datasets with the two control sets, we first calculated the frequency of occurrence for each residue type in each dataset and then expressed the composition of each amino acid in a given TF dataset as (TF-control)/(control). Thus, negative peaks indicate that TFs are depleted compared with the control in the indicated amino acids, and positive peaks indicate the reverse.

### *Statistical Analysis*

Analysis of variability in the percentage of proteins with predicted disorder was performed by bootstrap re-sampling. The 95% confidence intervals were calculated from the standard errors of 10 000 bootstrap iterations and are shown as error bars in Figures 3 thorough 6.

## RESULTS

*Dataset Characterization.* The workflow for dataset construction and nonredundancy preparation is shown in Figure 1. The overview of five datasets used in this study is present in Table 1 and is briefly outlined below. With the exception of PDBs25, datasets were filtered so that no two sequences have more than 25% identity using a serial sequence redundancy reduction, as described in the Material and Methods.

TF_SP&TRE_NR25 is a nonredundant representative dataset extracted from Swiss-Prot and TrEMBL. The nonredundant set contained 1819 sequences out of the initially retrieved set of 7195 sequences.

TF_SP_NR25 contains 1080 representative TF sequences. It is generally similar to TF_SP&TRE_NR25 after serial sequence redundancy reductions. The difference between TF_SP_NR25 and TF_SP&TRE_NR25 is that all entries in the former set were retrieved only from the Swiss-Prot database. This ensures that the feature table data containing domain information for every entry is well annotated and defined in Swiss-Prot format.

TF_NR25 was constructed on the basis of the availability of both sequence and feature table data (i.e., domain

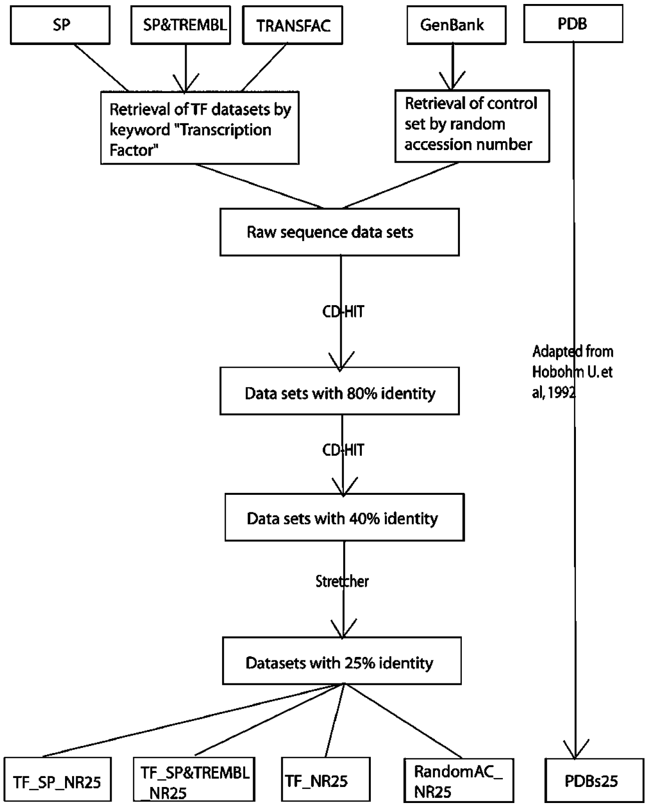
SP
SP&TREMBL
TRANSFAC
GenBank
PDB
Retrieval of TF datasets by keyword "Transcription Factor"
Retrieval of control set by random accession number
Raw sequence data sets
CD-HIT
Data sets with 80% identity
CD-HIT
Data sets with 40% identity
Stretcher
Datasets with 25% identity
Adapted from Hobohm U. et al, 1992
TF_SP_NR25
TF_SP&TREMBL_NR25
TF_NR25
RandomAC_NR25
PDBs25

FIGURE 1: Workflow for dataset construction and nonredundancy preparation.

information). The were 1186 TF sequences initially retrieved from TRANSFAC, which covers only eukaryotic cis-acting regulatory DNA elements and trans-acting factors. The final set contains 460 nonredundant sequences, whose homology is less than 25%.

PDBs25 (*26*) is a set of nonhomologous proteins, where no two entries have sequence identity higher than 25%, yet all structurally unique protein families are represented. These dataset contains 1583 sequences.

The Protein Data Bank (PDB) (*38*, *39*) contains more than 30 000 structures of proteins and protein complexes characterized by such methods as X-ray diffraction and nuclear magnetic resonance (NMR) spectroscopy. However, current information in the PDB is biased in the sense that it does not provide coverage of the whole of sequence/structure space. It has been shown that trans membrane, signaling, ID, and low complexity regions are significantly underrepresented in the PDB (*40*). To avoid this bias, we constructed a database called randomAC_NR25. We started with 4000 computer generated and randomized NCBI accession numbers, which corresponded to 2387 valid sequences retrieved from Genbank. After removing the sequence redundancy to less than 25%, 1930 sequences remained.

To better understand the degree of sequence redundancy in each preliminary dataset, a measure, which we called strip-off rate, was introduced (strip-off rate = (the number of sequences that are stripped off after reaching 25% identity)/(the original number of sequences)). Interestingly, the randomAC_NR25 dataset has a much lower strip-off rate (18.94%) compared to the three TF datasets. The strip-off rates for TF_SP&TRE_NR25, TF_SP_NR25, and TF_NR25 were 74.27%, 59.67%, and 61.21%, respectively (Table 1). The difference in sequence homology in the datasets is the major contributor to the difference in their strip-off rates.

*TF Amino Acid Compositional Specificity.* It has been shown that amino acid sequences encoding for ID proteins or regions are significantly different from those characteristic for ordered proteins on the basis of local amino acid composition, flexibility, hydropathy, charge, coordination number, and several other factors (*4*, *30*, *31*, *41*). A signature of a probable ID region is the amino acid compositional bias. This bias is characterized by a low content of so-called order-promoting residues, such as cross-linking Cys and bulky hydrophobic amino acids, such as Val, Leu, Ile, Met, Phe, Tyr and Trp, which would normally form the core of a folded globular protein, and a high proportion of particular polar and charged amino acids, such as Gln, Ser, Pro, Glu, Lys, and, on occasion, Gly and Ala, which are known as disorder-promoting residues (*23*, *42*).

The amino acid compositions of the TF sets were compared with those of the two control datasets (Figure 2) as described by Romero et al. 2001 (*23*). The fractional differences in composition for each amino acid between a given TF set and a control set was calculated as $(C_X - C_c)/C_c$, where $C_X$ is the averaged amino acid composition of a given TF dataset (TF_SP&TRE_NR25, TF_SP_NR25, or TF_NR25), and $C_c$ is the averaged amino acid composition in a control set (PDBs25 (C1, Figure 2A) or RandomAC_NR25 (C2, Figure 2B)). The amino acids in Figure 2 are arranged from the most rigid at the left to the most flexible at the right according to the scale of Vihinen (*43*). As the developers of this flexibility scale pointed out, the ranking does not reflect intrinsic flexibility; in which case, G would have the highest rank. Rather, the ranking depends on the degree to which a given side chain tends to be buried (low ranking) or exposed (high ranking) in the crystal structure of globular proteins. In agreement with the derivation of this scale, amino acids to the left have been shown to be order-promoting and those to the right disorder-promoting (*4*).

The amino acid compositions of TF_NR25, TF_SP_NR25, and TF_SP&TRE_NR25 are similar to each other but different from the two control sets. With the exception of histidine and glycine, all three TF datasets have compositional biases similar to ID proteins, relative to PDBs25 (*4*, *23*). Specifically, TFs are depleted in tryptophans, tyrosines, phenylalanines, isoleucines, and valines and enriched in serines, prolines, glutamines, and asparagines. Importantly, these results also reflect the specific signatures of TFs for DNA and protein binding. For example, histidines and cysteines are over-represented in TF datasets because nearly half of TFs contain multiple zinc fingers (*44*), and the most prevalent type of zinc fingers is the $C_2H_2$ zinc finger, which has been shown to play an important role in the overall recognition of DNA targets (*45*). Additionally, the high occurrence of proline and glutamine in these proteins also suggests that these residues may contribute to the conformational flexibility needed during the process of co-activators or repressors binding in transcriptional activation (*46*).

*Disorder Prediction on TF Datasets.* To test for a generalized prevalence of ID in transcriptional regulation, we predicted per-residue order−disorder using PONDR VL-XT and VSL1 to systematically analyze the abundance of ID in the three TF datasets (Figure 3). The results from both VL-XT (Figure 3A) and VSL1 (Figure 3B) predictors of ID

Table 1: Description of Five Nonredundant TF Datasets (NR25)

| dataset name | source | no. of entries in the raw dataset | strip-off rate[b] | no. of proteins | min. protein length (res.) | max. protein length (res.) |
|---|---|---|---|---|---|---|
| TF_SP&TRE_NR25[a] | Swiss-Prot and TrEMBL | 7195 | 74.27% | 1819 | 31 | 3859 |
| TF_SP_NR25[a] | Swiss-Prot | 2683 | 59.67% | 1080 | 53 | 6758 |
| TF_NR25[a] | TRANSFAC | 1186 | 61.21% | 460 | 109 | 3759 |
| RandomAC_NR25[a] | GenBank (randomized AC) | 2387 | 18.94% | 1930 | 31 | 5038 |
| PDBs25[a] | Hobohm et al. 1994 (*92*) | | | 1583 | 31 | 1235 |

[a] NR25 and s25: no two proteins have sequence similarity higher than 25% identical residues for aligned subsequences. [b] Strip-off rate: the number of sequences that have been stripped off after reaching 25% identity divided by the original number of sequences.

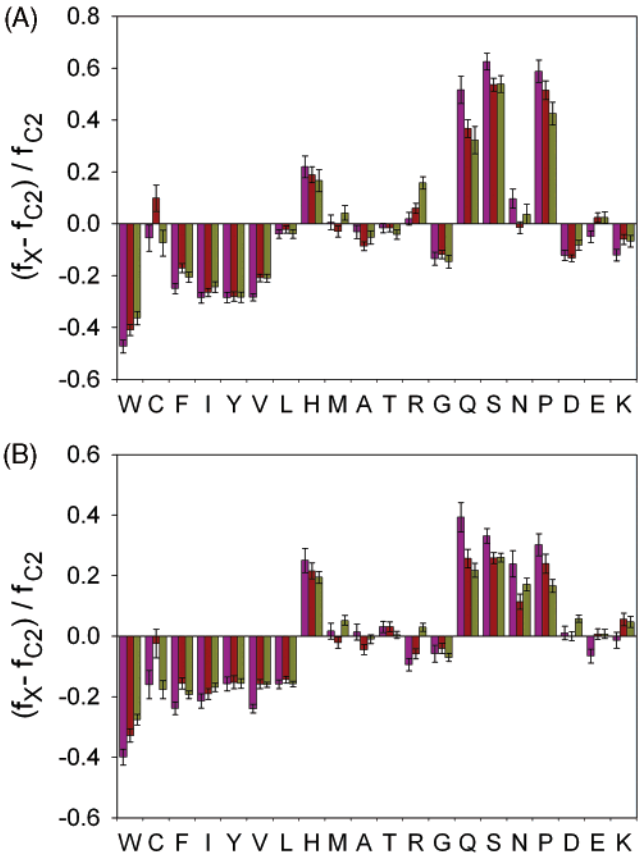

FIGURE 2: Composition profiling of TFs compared to PDBs25 (A) or RandomAC_NR25 (B). The bar for a given amino acid represents the fractional difference in composition between a set of TFs and a control set. The fractional difference is calculated as $(C_X - C_c)/C_c$, where $C_X$ is the composition of a given amino acid in a given TF database, and $C_c$ is the corresponding composition in a control set of proteins ($C_{c1}$ for PDBs25 and $C_{c2}$ for RandomAC_NR25). The residues are ordered by Vihinen's flexibility scale (*43*). Negative values indicate residues that the given TF set has less than the control, and positive indicates that it has more than the control. Error bars show one standard deviation.

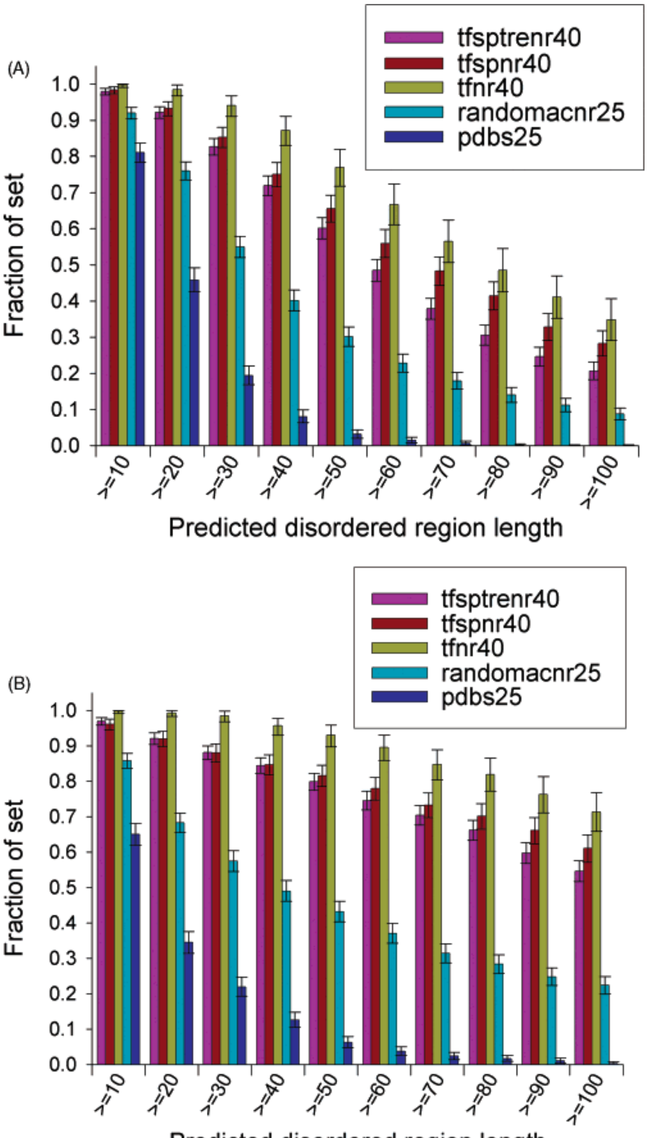

FIGURE 3: Summary of disorder predictions for the three TF datasets, TF_SP&TRE_NR25 (red bars), TF_SP_NR25 (pink bars), TF_NR25 (yellow bars), and two control sets, PDBs25 (blue bars) and RandomAC_NR25 (cyan bars), showing the fraction of proteins in each set with a predicted disordered region longer than that indicated on the horizontal axis. Predictions are summarized for (A) PONDR VL-XT and (B) PONDR VSL1, where the error bar shows the 99% bootstrap confidence interval.

are presented because the algorithms have relative strengths. Currently, PONDR VSL1 has been shown to be the most accurate predictor of order−disorder available, according to the results of the CASP6 experiments (*24*). However, observations indicate that VL-XT is much more sensitive to a local propensity for order, which has been used to identify binding regions within disordered proteins (*21*, *22*). That is, VL-XT may predict many disorder-to-order transition regions to be ordered.

To estimate the fraction of each set with significant disorder content, the fraction of proteins with predicted disordered regions of increasing length was calculated, which is a useful analysis because the rate of false predictions of disordered regions is inversely proportional to the length of the predicted disordered region (*1*). The fraction of each set predicted to contain disordered regions by both PONDR VL-XT (Figure 3A) and VSL1 (Figure 3B) followed the same ranking: TF_NR25 > TF_SP&TRE_NR25 > TF_SP_NR25 ≫ RandomAC_NR25 ≫ PDBs25. All three TF datasets, TF_NR25, TF_SP_NR25, and TF_SP&TRE_NR25, have significantly more disorder than PDBs25. The TF sets also contain significantly more disorder than the RandomAC_NR25 but to a lesser degree than PDBs25. The observed difference in the relative increase in disorder prediction is consistent with the nature of sequence selection and representation in the control sets, PDBs25 and RandomAC_NR25. Most proteins in PDB are ordered or partially ordered. In contrast, RandomAC_NR25 covers a much wider range of sequence diversity and is likely to contain many more disordered proteins.

The comparison among the three TF datasets suggested that TF_NR25 was more enriched in predicted disorder than the other two TF datasets, TF_SP_NR25 and TF_SP&TRE_NR25, where the difference between the later two sets was insignificant at most disordered region cutoff lengths. We believe that this is due to the fact that TF_NR25 was constructed using TRANSFAC, which is a dataset that covers only eukaryotic TFs. Recent studies suggest that the eukaryotic proteins have a higher percentage of ID than proteins from other kingdoms of life (*1*−*3*).

To gain further information on the abundance of ID in the TF datasets, their sequences were analyzed using two binary predictors of ID, cumulative distribution function (CDF) analysis (*1*) and the charge-hydropathy-plot (CH-plot) (*25*). Both of these methods perform a binary classification of whole proteins as either mostly disordered or mostly ordered, where mostly ordered indicates proteins that contain more ordered residues than disordered residues, and mostly disordered indicates proteins that contain more disordered residues than ordered residues (*3*). CH-plots and CDF analysis predict different extents of disorder in many datasets, and it has been proposed that these discrepancies may be physically interpretable (*3*). In general, CH-plot analysis is predisposed to discriminate proteins with substantial amounts of extended disorder (random coils and pre-molten globules) from proteins with globular conformations (molten globule-like and rigid well-structured proteins), whereas PONDR-based CDF analysis may discriminate all disordered conformations including molten globules from rigid well-folded proteins. Therefore, this analysis may be useful in estimating the types of disorder present in TFs.

CDF analysis summarizes the per-residue disorder predictions by plotting PONDR VL-XT scores against their cumulative frequency, which allows ordered and disordered proteins to be distinguished on the basis of the distribution of prediction scores. In this case, order−disorder classification is based on whether a CDF curve is above or below a majority of boundary points (*3*). An example of the CDF plots for randomly selected proteins is shown in Figure 4A, and a summary of the fraction of wholly disordered proteins in each set is shown in Figure 4B. CDF analysis predicts that ≥70% proteins in the TF datasets are wholly disordered, which is greater than the number of wholly disordered proteins in the RandomAC_NR25 and PDBs25 datasets (28.81% and 17.75%, respectively). This agrees with the estimates of disorder based on the length of the predicted disordered region.

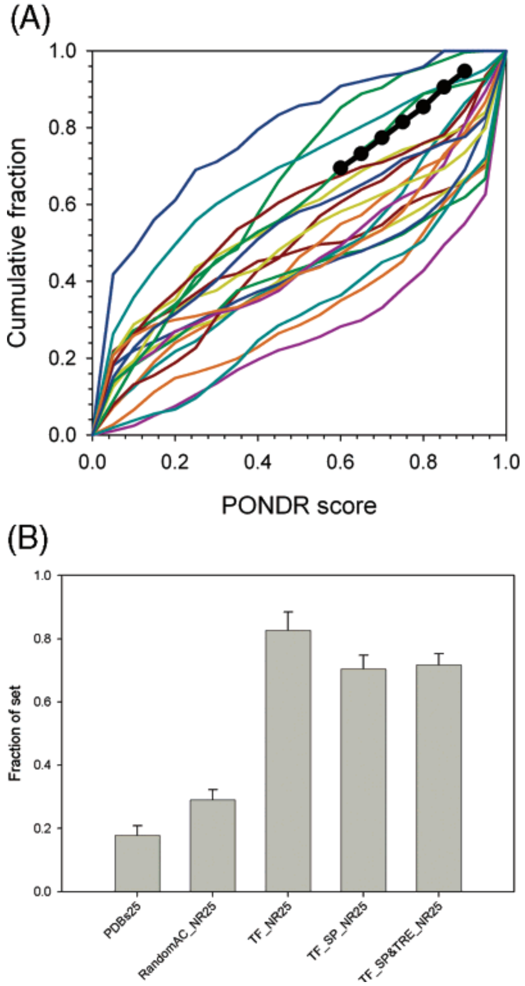

FIGURE 4: PONDR CDF analysis of whole protein order−disorder. (A) Sample plot of the CDF curves for randomly selected proteins from the TF_SP_NR25 set, where proteins are shown by the colored lines and the order−disorder boundary is shown by the black line. (B) Fraction of each of the five sets predicted to be wholly disordered by CDF analysis. Error bars show the 99% bootstrap confidence interval.

Ordered and disordered proteins plotted in CH space can be separated to a significant degree by a linear boundary, with proteins located above this boundary line (i.e., proteins with a relatively high net charge and/or a low hydropathy) being disordered and the proteins below the boundary line (i.e., proteins with a relatively low net charge and/or a high hydropathy) being ordered (*25*). An example of the CH-plot of randomly selected proteins is shown in Figure 5A. The percentage of wholly disordered proteins predicted by CH-plot analysis in the three TF datasets is between 28.70% and 23.80% (Figure 5B). These values, being smaller than that predicted by CDF analysis, are still substantially higher in comparison with RandomAC_NR25 and PDBs25 (10.36% and 16.99%, respectively).

The results of CDF and CH-plot analyses show sizable discrepancies, and in four datasets (TF_NR25, TF_SP_NR25, TF_SP&TRE_NR25, and RandomAC_NR25), the level of disorder predicted by CDF was 2.78- to 2.94-

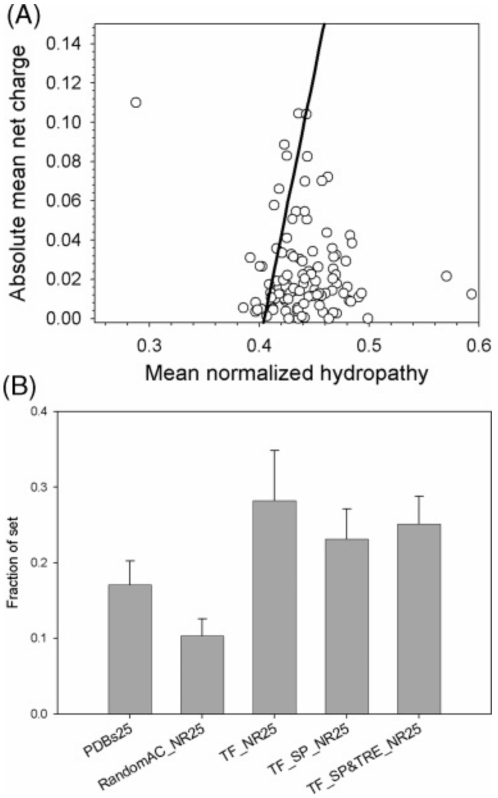

FIGURE 5: CH-plot analysis of whole protein order−disorder. (A) Sample plot of randomly selected proteins from the TF_SP_NR25 set in CH space, where proteins are shown by the cycles and the order−disorder boundary is shown by the black line. (B) Fraction of each of the five sets predicted to be wholly disordered by CH plots. Error bars show the 99% bootstrap confidence interval.

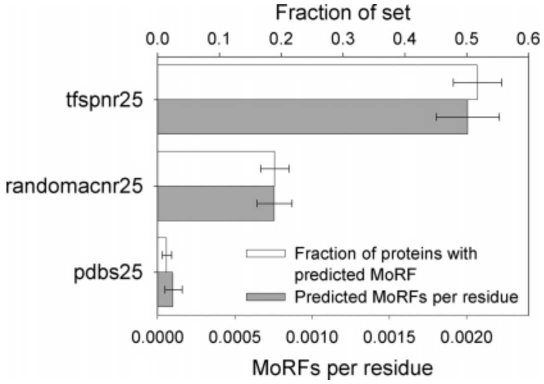

FIGURE 6: α-MoRF predictions for the TF_SP_NR25 set and the two control sets. Both the faction of proteins containing predicted α-MoRFs (white bars) and the predicted α-MoRFs/residue in each set (gray bars) are shown. Error bars show the 99% bootstrap confidence interval.

fold higher than that predicted by CH plots. The difference between these two methods is consistent with other results (*3*).

We also noted that ∼17% of proteins in PDBs25 were predicted to be wholly disordered by both CDF and CH-plot analyses. Wholly disordered proteins are not expected to crystallize. At first glance, this implies that the PDBs25 proteins predicted to be wholly disordered might represent prediction errors. However, many ID proteins become ordered upon binding to partners. Such proteins can appear in the PDB as ordered because the complex, not the individual protein, has been crystallized. Alternatively, ID proteins can fold (and rigidify) as a result of the interaction with the co-solutes used in crystallization. Some ID proteins or protein regions can gain structure by being involved in crystal contacts.

In addition to disorder predictions, we also predicted α-MoRFs for the TF_SP_NR25, randomAC_NR25, and PDBs25 datasets. These predictions indicate the presence of short regions within longer regions of disorder that are responsible for the recognition of protein-binding partners. These short regions are disordered in the absence of binding partners and undergo a disorder-to-order transition upon binding to molecular partners, thereby forming a stable, partially helical structure. The results of the α-MoRF prediction are shown in Figure 6. PDBs25 was the negative control for the development of the α-MoRF predictor and has a very low level of α-MoRF predictions by design. The level of α-MoRFs predicted in the RandomAC_NR25 set, ∼19% of proteins or ∼7.5 α-MoRFs per 10,000 residues, represents the frequency of α-MoRFs across all types of proteins. The level of α-MoRFs prediction in TF_SP_NR25 was much higher, ∼50% of proteins or ∼2 α-MoRFs/1000 residues, which suggests that α-MoRFs are very common in TFs. This result agrees with previous results, which implicated α-MoRFs in proteins with transcriptional regulation activity (*21*).

*Disorder in TF Domains and Subdomains.* To gain insight into the association between TF functions (DNA-binding and transcriptional regulation) and ID, we systematically dissected TFs into various annotated domains in one of the TF datasets (TF_SP_NR25) and analyzed the coupling of the region of disorder or order with established function. The domain annotation was extracted from the TF_SP_NR25 dataset in Swiss-Prot format. The disorder predictions were calculated, and grouped on the basis of the key name combined with the description in the corresponding feature table as described in the Materials and Methods section. Predictions of ID for both DNA-binding domains (DBDs) (Table 2) and activation domains (Table 3) are presented. One may notice that some of the DBD names are different from the 53 DNA-binding domains in the Pfam models that have been reported by several groups (*47*, *48*). This is due to the difference in methods used for domain annotations and classifications in Swiss-Prot and Pfam. Below, the results for basic domains, helix-turn-helix, AT hook, zinc fingers and their linkers, and activation domains are given in detail.

The basic domain is one of the most frequent motifs among the TF_SP_NR25 DBDs. Basic domains range between 4 and 33 residues long, where the mean length is 18.1 residues. They are characterized by a high percentage of predicted disorder, 96.6%, which suggests that basic domains in TFs are in general highly unstructured. This computational analysis is strongly supported by a wealth of experimental data. In the early 90s, it was demonstrated that the basic regions of the *bzip* proteins Fos and Jun (*49*), C/EBP (*50*), and GCN4 (*51*) are unfolded in their unbound states. The basic motifs are characterized by a large excess of positive charges, which prevents them from being structured when free in solution, but fold and form α-helical structure when interacting with DNA (*52*). Usually, the basic domain appears in close proximity to a dimerization domain, such as a leucine zipper. After dimerization, the highly

Table 2: Order−Disorder Region in TF DBD

| motif name | no. of motifs in the dataset | average motif length (res.) | VLXT | VSL-1 |
|---|---|---|---|---|
| | | | overall disorder | overall disorder |
| DNA_BIND: A.T hook | 19 | 12.21 | 99.14% | 100.00% |
| DNA_BIND: AP2/ERF | 12 | 58.42 | 44.08% | 12.13% |
| DNA_BIND: basic motif | 98 | 18.12 | 96.62% | 96.06% |
| DNA_BIND: CUT | 7 | 87.86 | 43.41% | 35.77% |
| DNA_BIND: ETS | 14 | 82 | 21.25% | 2.87% |
| DNA_BIND: fork-head | 20 | 92.75 | 17.25% | 17.20% |
| DNA_BIND: H-T-H motif | 19 | 20.05 | 37.80% | 6.04% |
| DNA_BIND: HMG box | 39 | 70.33 | 46.08% | 54.14% |
| DNA_BIND: tryptophan cluster | 5 | 102.2 | 27.98% | 20.35% |
| DNA_BIND: homeobox | 91 | 60.73 | 46.80% | 35.72% |
| DNA_BIND: Myb | 15 | 50.73 | 48.49% | 6.40% |
| DNA_BIND: NR-type | 8 | 73.38 | 39.35% | 24.36% |
| DNA_BIND: T-box | 13 | 164.69 | 12.19% | 10.09% |
| DNA_BIND: WRKY | 66 | 66.61 | 31.23% | 30.46% |
| DNA_BIND: Zn(2)−Cys(6) | 13 | 28.62 | 48.12% | 86.02% |
| ZN_FING: C2H2-type | 1016 | 23.69 | 9.50% | 30.01% |
| ZN_FING: C4-type | 21 | 24.1 | 26.28% | 20.55% |
| ZN_FING: B box-type | 13 | 45.62 | 9.61% | 30.86% |
| ZN_FING: C2HC-type | 19 | 24.42 | 7.76% | 40.73% |
| ZN_FING: Dof-type | 7 | 55 | 29.09% | 35.58% |
| ZN_FING: GATA-type | 21 | 25.1 | 3.23% | 34.91% |
| ZN_FING: MYM-type | 9 | 40.67 | 0.55% | 28.69% |
| ZN_FING: PHD-type | 29 | 52.97 | 8.20% | 18.75% |
| ZN_FING: RING-type | 17 | 44.24 | 12.10% | 19.55% |
| ZN_FING: Zn-ribbon | 9 | 26.44 | 0.42% | 9.05% |
| average | | | 30.66% | 32.25% |

Table 3: $C_2H_2$ Linker Information (VLXT)

| length (res.[a]) | no. of linkers | total no. of res. | no. of disordered res. | overall disorder |
|---|---|---|---|---|
| linker = TGEKP | 142 | 710 | 49 | 6.90% |
| linker = 5 | 518 | 2590 | 207 | 7.99% |
| 1≤ linker ≤ 10 | 659 | 3348 | 282 | 8.42% |
| 11 ≤ linker ≤ 50 | 96 | 2658 | 1179 | 44.36% |
| 51 ≤ linker ≤ 100 | 47 | 3464 | 2153 | 62.15% |
| linker > 100 | 76 | 18600 | 12180 | 65.48% |

[a] res.: residue.

disordered basic region is believed to mediate sequence-specific DNA binding via an induced-fit recognition of DNA (*53*, *54*).

Similar to the basic domains, A−T hook domains are predicted to be nearly completely disordered (Table 2). The AT hook is a small DBD that was first described in the high mobility group nonhistone chromosomal protein HMGI/Y (*55*−*57*). This motif preferentially binds to the narrow minor groove of stretches of AT-rich sequences and participates in a wide variety of cellular processes including the regulation of inducible gene transcription (*55*, *58*). HMG-type and T-box are classified to be in the group of $\beta$-scaffold domains with minor groove contacts according to this specific DNA-binding feature. Various physical studies, including NMR spectroscopy, have demonstrated that as free molecules in solution the HMGI/Y proteins have no detectable secondary or tertiary structure (*59*).

Another important type of DBD is the helix-turn-helix domain. It is one of the most striking substructures in this superfamily: homeobox, forked/winged helix, tryptophan clusters, and mybdomains. The analysis of disorder predictions on this class of protein domains demonstrates that the propensity for ID follows the ranking Myb > homeobox > tryptophan clusters > forked/winged helix (Table 2), with the percentage of overall disorder (VL-XT predicted) being 48.49, 46.8, 27.98, and 17.25%, respectively. The helix-turn-helix domain is characterized by two α-helices, which make extensive contacts with DNA and are joined by a short turn. The second helix binds to DNA via a number of hydrogen bonds and hydrophobic interactions, which occur between specific side chains and the exposed bases and thymine methyl groups within the major groove of the DNA. The first helix helps to stabilize the structure (*60*).

Zinc-finger motifs were originally identified as DNA-binding structures in the RNA polymerase III TF TFIIIA, which binds to the internal control region of the 5S RNA gene (*61*). Wright and co-workers determined the first structure of an isolated zinc-finger domain by solution NMR (*62*). At least two types of zinc fingers (the classic zinc finger proteins and the steroid receptor (i.e., glucocorticoid receptor)) have also been found in TFs, which mediate the transcription mediated by PolII. Five classes, $C_2H_2$, $C_4$, DM, GCM, and WRKY, currently populate the super class of zinc-coordinating domains (*48*). Table 2 suggests that the $C_2H_2$ zinc-finger domain, which is the most abundant DBD in the TF_SP_NR25 dataset, is one of the most highly ordered DNA-binding protein motifs. Its overall percentage of predicted disorder is as low as 9.5%. The $C_2H_2$ zinc-finger motif is prevalent in the mammalian TFs and the TFs of other eukaryotes (*44*). It consists of an average of 24 amino acids (the shortest is 12, and the longest is 38 residues) with 2 cysteine and two histidine residues that bind the zinc ion. This $Zn^{2+}$ coordination folds the relatively short polypeptide into a compact domain. The highly ordered zinc motif provides a rigid and stable structure for docking arrangement and base recognition to insert its α-helix into the major groove of DNA (*45*).

Many of the zinc-finger motifs act as independently folded globular domains ($\beta\beta\alpha$) that are separated by flexible linker regions. The linker region is an important structural element that helps control the spacing of the fingers along the DNA site and, thus, contributes to the coordination of TF activity. Also, the properties of the linker region can have a large impact on the affinity of connected DBDs for its target DNA (*63*). To better understand the role of ID in these linkers, we used the linker region connecting two $C_2H_2$ zinc fingers in the TF_SP_NR25 set to predict its ID propensity and analyze the peculiarities of amino acid composition in this region. The results presented in Table 3 indicate that linkers separating two $C_2H_2$ domains vary greatly in length and composition.

More than half of the linkers have five residues between the final histidine of one finger and the first conserved aromatic amino acid of the next finger. Over one-quarter of these five-residue-long linkers have a consensus sequence of TGEKP. It has been shown by NMR that the TGEKP linker is flexible in the free protein but becomes more rigid upon DNA binding (*64*−*66*). However, these short linkers are predicted to be highly ordered (Table 3). This prediction is likely to be an artifact of the windowing procedure used for PONDR predictions. For example, PONDR VL-XT uses an effective window size of 29 residues of which the five-residue linker contributes only ∼6% of the window. Combined with the extreme order predictions for neighboring $C_2H_2$ domains, this causes short linkers to be predicted to

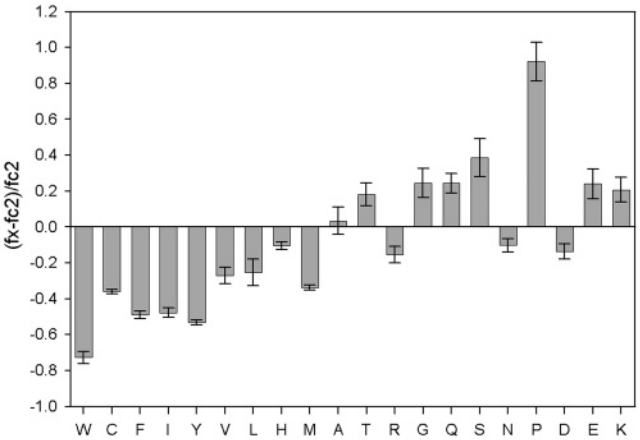

FIGURE 7: Composition profiling of $C_2H_2$ zinc-finger linkers compared to RandomAC_NR25. The bar for a given amino acid represents the fractional difference in composition between a set of linkers and a control set. The fractional difference is calculated as ($C_{\mathrm{Linker}} - C_{\mathrm{control}}$)/$C_{\mathrm{control}}$, where $C_{\mathrm{linker}}$ is the composition of a given amino acid in a database of linkers, and $C_{\mathrm{control}}$ is the corresponding composition in the RandomAC_NR25 dataset. The residues are ordered by Vihinen's flexibility scale (Vihinen, M.1987). Negative values indicate residues that are under-represented in a linker dataset, whereas positive values correspond to the over-represented.

be ordered. In support of this argument, we found that the magnitude of predicted disorder increases with the length of the linker. The overall percentage of disorder is as high as 65% when the linker length reaches 100 residues.

The analysis of amino acid compositions also demonstrates that the linker regions not only have high occurrences of the five residues of the consensus sequence, TGEKP, but also show increased prevalence of some polar, uncharged amino acids and are substantially depleted in many of the order-promoting amino acids (Figure 7). Conclusive evidence that the linker length and composition can influence both binding specificity and affinity independent of the DNA-binding subdomains has come from several recently published studies (*63*, *67*−*69*).

TF activation domains consist of 30−100 amino acids and are distinct from the DBDs. There are three different primary sequence motifs that have been identified: acidic, glutamine-rich, and proline-rich domains. Deletion analyses of numerous TFs from mammals and *Drosophila* have identified several other classes that are rich in serine, threonine, or other amino acid residues containing hydroxyl groups. Also, some strong activation domains that are not particularly rich in any specific amino acid have also been identified. Additionally, a few repression domains have been identified; the best characterized of which is the alanine-rich domain. Poor conservation made it difficult to thoroughly analyze activation domains, and we only used the five different primary sequence motifs mentioned above for our study. PONDR predicts a high degree of disorder in these activation domains (Table 4), ranging from 73 to 94%. There is no publicly available solved structure for any trans activation domains, whereas DNA-binding domains for many structures have been determined. The lack of known structure supports our finding that most transcriptional activation domains are likely to be either completely or mostly unstructured.

Table 4: Disorder Analysis of TF Activation Domains

| domain name | average domain length (res.[a]) | no. of disordered res. | overall disorder | no. of domains in the dataset |
|---|---|---|---|---|
| acidic | 40 | 31 | 77.48% | 81 |
| alanine_rich | 20 | 19 | 94.52% | 11 |
| glutamine_rich | 46 | 34 | 73.33% | 55 |
| glycine_rich | 29 | 26 | 88.98% | 16 |
| proline_rich | 61 | 48 | 77.65% | 30 |
| serine_rich | 36 | 33 | 93.13% | 13 |
| average | | | 84.18% | |

[a] res.: residues.

*Comparison of PONDR Predictions and Known Structure.* It is informative to directly compare plots of per-residue order prediction and known structural features of a given protein. Such comparisons can reveal features of the proteins in question that are not obvious from the isolated information. To this end, Figure 8 was generated to compare the PONDR predictions and known structural information for three proteins from humans: TFIIA (α and $\beta$ subunits), TFIIB, and TATA box-binding protein (TBBP). These proteins were selected from a list of proteins in the TF_SP_NR25 dataset with multiple chains in the PDB, generated by filtering the SwissProt-PDB mapping available from SwissProt and selected on the basis of their well-known biological role in polII-dependent transcription. The Figure shows the PONDR VL-XT plots for the three proteins, a complex between TBBP and TFIIA, a complex between TBBP and TFIIB, and the zinc ribbon domain of TFIIB. Also indicated are the residues in contact in the respective TBBP complexes.

The most striking feature of this Figure is the good agreement of predictions of order with regions of known structure in the PDB. TFIIA contains extensive regions of predicted disorder, but the α and $\beta$ subunits correspond nearly exactly to predicted ordered regions. The same is true for the zinc ribbon of TFIIB. The DBDs of TBBP and TFIIB also correspond to regions predicted to be ordered; however, these sequences have some predicted disorder within the regions of known structures. An examination of the contact residues between TFIIB and TBBP shows that some of these disordered regions correspond to the interface in this complex. In addition, other contact residues correspond to peaks in the disordered prediction, albeit these peaks are below the usual prediction threshold. These correspondences suggest that interaction regions have a strong propensity for disorder and may be disordered regions that become ordered upon binding.

Another prominent feature of Figure 8 is the long predictions of disorder in TFIIA, which are interrupted by two sharp short predictions of order. This feature has been noted in the context of other proteins by two groups, called indications of binding regions (*22*) or regions of intrinsic structural preference (*37*), and form the basis for an α-MoRF predictor (*21*). Though not predicted to be α-MoRFs, these regions may correspond to protein or specific DNA recognition sites within a long disordered region. Structural characterization of these regions should help to elucidate the mechanisms of TF function.

*Top 15 Predictions of Disordered TFs.* To provide illustrative examples, 15 TFs with the highest rankings of ID were selected. To avoid ortholog redundancy, these proteins were selected from a single organism, *Homo sapiens*,

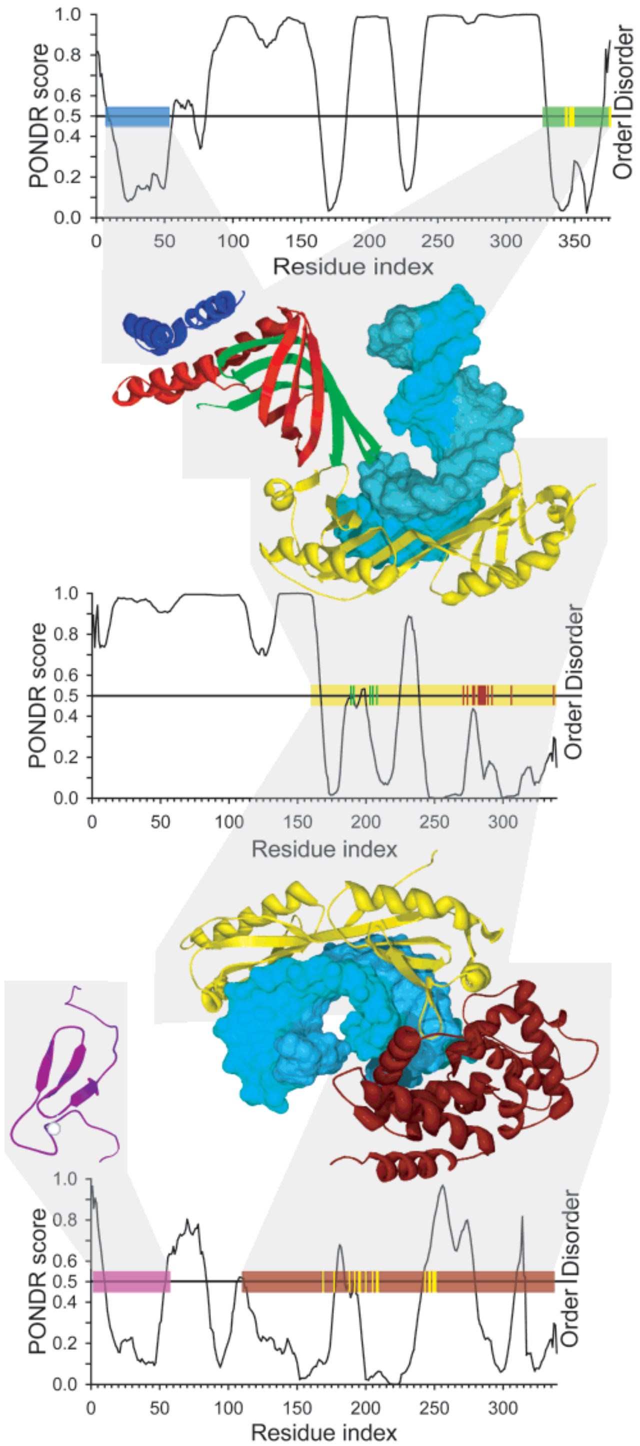

FIGURE 8: PONDRing TFIIA, TFIIB, and TATA box-binding protein. The correspondence of PONDR predictions and regions of known structure are shown. Three PDB structures are shown; 1NVP (top), 1C9B (bottom right), and 1DL6 (bottom left), where each chain in the ribbon and molecular surface representations are color coded. TATA box-binding protein (yellow), TFIIA-α (blue), TFIIA-β (green), TFIIA-γ (red, PONDR prediction not shown), TFIIB core domain (maroon), and TFIIB zinc-ribbon domain (purple). These color codes are also used for the bars in the three PONDR plots, (top) TFIIA, (middle) TATA box-binding protein, and (bottom) TFIIB, to indicate the positions of the regions of known structure in the context of the PONDR predictions. Drawn over these bars, hash marks show the residues in contact with other chains, where the color of the hash mark corresponds to the color code of the partner chain.

Table 5: Top 15 Human TFs from TF_SP_NR25 with >80% of the Residues Predicted

| gene code | TF name | length | disorder res.[a] | overall disorder | LDR[b] | exp. confirmed |
|---|---|---|---|---|---|---|
| HMGN1; HMG14 | nonhistone chromosomal protein HMG-14 | 99 | 99 | 100.00% | 99 | yes |
| HMGA2; HMGIC | high mobility group protein HMGI-C | 109 | 109 | 100.00% | 109 | yes |
| NFYA | CCAAT-binding transcription factor subunit B | 347 | 334 | 96.25% | 262 | |
| SP4 | transcription factor Sp4 | 784 | 706 | 90.05% | 440 | |
| TFE3 | transcription factor E3 | 743 | 656 | 88.29% | 169 | |
| CREB1 | cAMP response element binding protein (CREB) | 341 | 299 | 87.68% | 212 | yes |
| SOX15 | SOX-15 protein | 233 | 201 | 86.27% | 105 | yes |
| RREB1 | RAS-responsive element binding protein 1 | 755 | 648 | 85.83% | 239 | |
| SOX3 | transcription factor SOX-3 | 446 | 374 | 83.86% | 115 | yes |
| MAFF | transcription factor MafF | 164 | 137 | 83.54% | 69 | |
| CEBPG | CCAAT/enhancer binding protein gamma | 150 | 125 | 83.33% | 106 | |
| SRF | serum response factor (SRF) | 508 | 423 | 83.27% | 165 | |
| SP2 | transcription factor Sp2 | 606 | 497 | 82.01% | 243 | |
| DBPA | DNA-binding protein A | 372 | 303 | 81.45% | 115 | |
| BATF | ATF-like basic leucine zipper transcription factor B-ATF | 125 | 101 | 80.80% | 68 | |

[a] res.: residues. [b] LDR: longest disordered regions.

from TF_SP_NR25 (Table 5). As a consequence of ranking TFs on the basis of overall percentage of disordered residues predicted by PONDR (above 80%), these proteins represent extremes of long predicted disordered regions and high disorder prediction scores. It was surprising to find that four ID proteins (HMGI-14, HMCI-C, SOX-15, and SOX-3) among these Top 15 TFs belong to one superfamily, the high mobility group (HMG). HMG is composed of three different families that have recently been renamed HMGA (also known as HMGI/Y), HMGB (also known as HMG-1 and -2), and HMBN (also known as HMV-14 and -17) (*70*). HMG proteins are the founding members of a new class of regulatory elements called architectural TFs that participate in a wide variety of cellular processes, including the regulation of inducible gene transcription, integration of retroviruses into chromosomes, the induction of neoplastic transformation, and the promotion of metastatic progression of cancer cells (*56*). To illustrate the important and central role of these TFs in the various biological processes and to identify potential signaling pathways that they may participate in, one of the HMG family members, HMGA, was selected for analysis with PathwayAssist v3.0 (www.ariadnegenomics.com). PathwayAssist is used to visualize and explore the biological pathways, gene regulation networks, and protein−protein interactions. The ResNet, supplied with PathwayAssist, is a molecular interaction and pathway database which contains more than 500 000 functional links for more than 50 000 proteins, extracted from more than

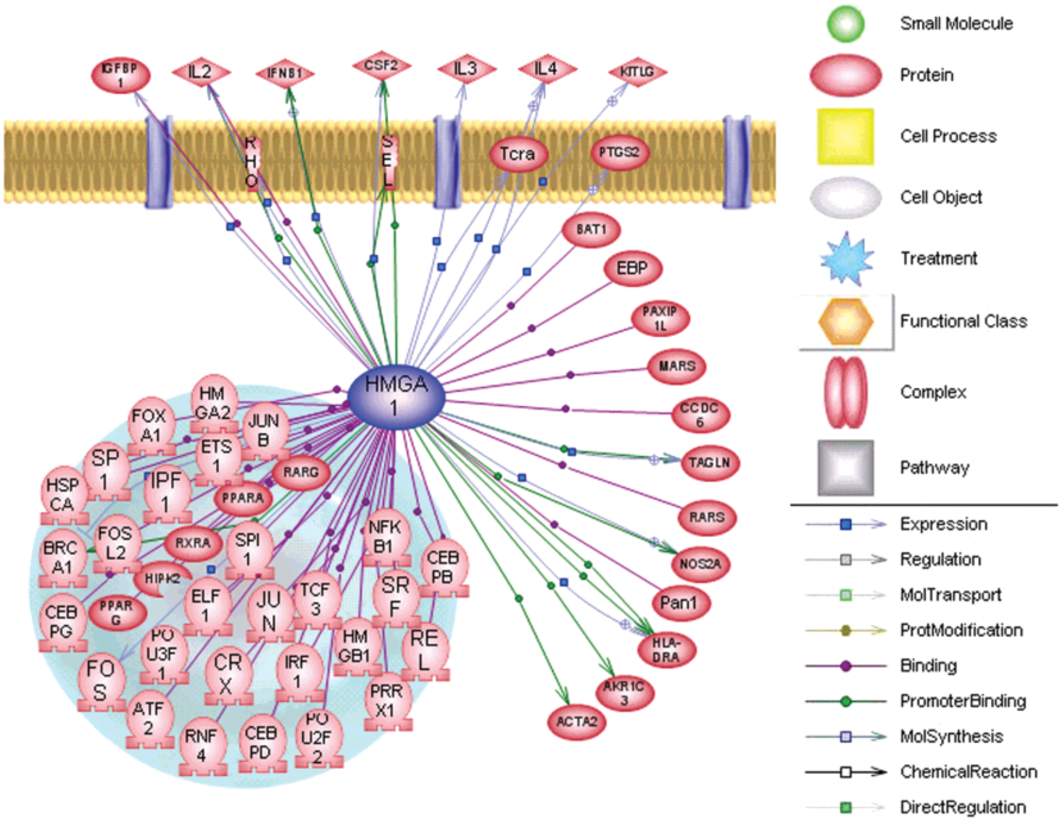

FIGURE 9: Analysis of the molecular interactions of HMGA1 using PathwayAssist. A simplified network associated with HMGA1 is shown. HMGA1 protein acts as a 'hub' of nuclear function and interacts with at least 18 TFs in the nucleus.

Table 6: Analysis of TF Disorder in Different Species

| species name | average length (res.) | overall ID | LDR[a] ≥30 | LDR ≥40 | LDR ≥50 | no. of proteins |
|---|---|---|---|---|---|---|
| human | 652.85 | 49.90% | 91.50% | 83.28% | 77.71% | 341 |
| mouse | 673.07 | 50.15% | 86.03% | 82.35% | 74.26% | 136 |
| yeast | 565.19 | 39.43% | 85.07% | 73.88% | 59.70% | 134 |
| rat | 430.56 | 47.94% | 90.98% | 73.77% | 59.02% | 122 |
| mouse-ear cress | 418.42 | 46.64% | 90.65% | 71.96% | 56.07% | 107 |
| fruit fly | 666.13 | 54.30% | 95.65% | 91.30% | 85.51% | 69 |
| *Bacillus subtilis* | 284.85 | 23.07% | 40.00% | 25.00% | 15.00% | 40 |
| fission yeast | 474.73 | 40.99% | 78.79% | 66.67% | 60.61% | 33 |
| *C. elegans* | 488.04 | 41.04% | 82.61% | 60.87% | 56.52% | 23 |
| zebra fish | 370.61 | 52.60% | 83.33% | 72.22% | 61.11% | 18 |
| *Escherichia coli* | 343.25 | 30.79% | 62.50% | 50.00% | 12.50% | 16 |

[a] LDR: longest disorder region.

5 000 000 Medline abstracts and full-length articles (ResNet update Q4, 2004). The result as shown in Figure 9 illustrates that HMGA proteins specifically interact with more than 20 other proteins, most of which are TFs.

*TF Disorder in Different Species.* Disorder predictions for 11 model organisms with proteins in the TF_SP_NR25 dataset are shown in Table 6. From the predictions summarized here, it is apparent that TFs from eukaryotes are likely to be far more disordered than bacterial TFs. The percentage of sequences predicted to contain long disordered segments (≥50 consecutive disordered residues) in eukaryotes ranged from 56% for Mouse-ear cress to 77% for human, whereas in prokaryotes it ranged from 12.5 to 15%. The generality of this conclusion might be questioned by the fact that the average length of the eukaryotic sequences is significantly greater than that of bacterial proteins. However, the overall percentage of disorder, a length-independent parameter, was also shown to be essentially higher in TFs from eukaryotes.

## DISCUSSION

*Potential Data Set Bias.* There are more than 30 000 structures of proteins and protein complexes in the Protein Data Bank (PDB) (*38*). A comparison of the nonredundant subset of PDB with several complete genomes revealed that proteins encoded by the genomes are significantly different from those in the PDB with respect to sequence length, amino acid composition, and predicted secondary structure composition (*71*). Thus, information currently present in the PDB is highly biased in the sense that it does not provide complete coverage of the whole of sequence/structure space (*72*). It has been established that trans membrane, signal, disordered, and low complexity regions are significantly under-represented in PDB, whereas disulfide bonds, metal binding sites, and sites involved in enzyme activity are over-represented (*40*). Additionally, hydroxylation and phosphorylation and posttranslational modification sites were found to be under-represented, whereas acetylation sites were significantly over-represented (*40*). PDBs25 should therefore be viewed as representative of known structures, rather than representative of all possible structures. (Theoretically, it contains one member of each protein family (*73*).)

In contrast to PDBs25, the RandomAC_NR25 dataset, which we constructed as an additional control set, does not possess the same degree of bias. For example, the fact that the RandomAC_NR25 dataset has a much lower strip-off rate (18.9%) compared to the strip-off rates of the three TF datasets (the rate for TF_SP&TRE_NR25, TF_SP_NR25, and TF_NR25 are 74.3, 59.7, and 61.2%, respectively) indicates that the raw dataset used to derive the RandomAC_NR25 was much less redundant. The extent of sequence homology in the different datasets explains the difference. Because the RandomAC_NR25 set was retrieved via the randomized choice of accession numbers, it contains well-represented broad sequence diversity and likely reflects the natural sequence/structure complexity and, thus, better covers the sequence/structure space. However, all sequences deposited in the three raw TF datasets are TFs. Some of the entries in the raw TF databases have probably been derived from the same superfamilies and have similar highly conserved motifs, such as the zinc finger, leucine zipper, homeobox, and so forth.

*ID in TFs.* Our data are also consistent with the conclusion that the degree of disorder in eukaryotic TFs is significantly higher than in prokaryotic proteins. Eukaryotes have a well-developed gene transcription system, which probably requires a great deal of flexibility and plasticity. The intrinsically disordered TFs or partially unstructured regions can offer significant advantages in response to different molecular targets, allowing one protein to interact with multiple cellular partners and allowing fine control over binding affinity.

Our data also indicates that there are two distinct types of DBDs in TFs. Many types of DBDs are likely to be well structured and specifically recognize DNA on the basis of the molecular surfaces they present to the environment. Other DBDs, such as basic domains or AT hooks, are likely to be highly unstructured in isolation and presumably undergo a disorder-to-order transition upon binding to specific DNA. The functional requirements that have selected these two seemingly orthogonal modes of DNA binding activity are in general, not well understood. One possible explanation is that binding and folding forms structured binding sites, which do not exist in the unfolded state. This would preclude interaction with some partners prior to binding to specific DNA and enforce a sequential assembly of TF complexes.

The predicted high abundance of intrinsically disordered transcription activation domains provides strong support for a physiological role of coupled folding and binding processes in transcriptional activation. The inherent flexibility of transactivation domains provides their local and global structure with a unique possibility to be modified in response to interaction with different molecular targets, allowing one protein to interact with multiple cellular partners and ensuring fine control over binding affinity.

The difference in the magnitude of ID predicted by CH plots and CDF may hold clues to the structural state of some TFs. This discrepancy has been reported before, and it has been proposed that it can be physically interpretable (*3*). Specifically, the CH plot is a linear classifier that takes into account only two parameters of the particular sequence, charge and hydrophobicity (*25*), whereas CDF analysis is dependent upon the output of the PONDR VL-XT predictor, a nonlinear neural network classifier, which was trained to distinguish order and disorder on the basis of a significantly larger feature space that explicitly includes net charge and hydropathy (*29*−*31*). Thus, the CH space represents a subset of PONDR VL-XT feature space (*3*). It may be that CH-plot analysis is predisposed to discriminate proteins with substantial amounts of extended disorder (random coils and pre-molten globules) from proteins with globular conformations (molten globule-like and rigid well-structured proteins). However, the PONDR-based CDF analysis may discriminate all disordered conformations including molten globules from rigid well-folded proteins. However, this means that some ID TFs are possibly extended, whereas others may possess molten globule-like properties. Regardless of the differing degree of predicted disorder, both classifiers predicted TF datasets to contain a high portion of wholly disordered proteins. This finding confirms the trends discovered by several individual case studies on a small number of TFs analyzed to date (*9*).

*ID and TF Function.* We have demonstrated that the three TF datasets contain a significantly greater number of ID proteins than the two control sets. This implies that ID may play a critical role in the primary functions of TFs, which are molecular recognition, DNA binding, and transcriptional regulation. This hypothesis is in a good agreement with recent experimental findings, some of which are outlined below.

Although the occurrence of unstructured regions of significant size (>50 residues) is surprisingly common in functional proteins, the crucial role of intrinsically disordered regions in the activity of TFs has only recently been recognized. The last two decades have witnessed a tremendous expansion in our knowledge of TFs and the way in which they are employed by eukaryotic cells to control gene activity. Ample evidence has accumulated showing that eukaryotic TFs contain a variety of structural motifs that interact with specific DNA sequences (*74*) and are involved in activating transcription. For example, on the basis of the analysis of the binding of multiple zinc fingers to cognate DNA, a snap-lock model has been recently introduced (*75*). According to this model, $C_2H_2$-type zinc-finger domains consist of well-folded modules connected by highly conserved linker sequences that are mobile and unstructured in the absence of the cognate DNA. Importantly, our findings are in good agreement with these results (see Tables 2 and 3), which suggest that in general $C_2H_2$ zinc-finger domains are highly ordered, whereas linkers possess significant ID. NMR analysis revealed that upon binding to the correct DNA sequence the linker becomes highly structured and locks adjacent fingers in the correct orientations in the major groove (*75*). Furthermore, it has been shown that any alterations to this linker (e.g., alternative splicing leads to the insertion of the tripeptide KTS in the linker between the third and fourth $C_2H_2$ zinc fingers of the Wilms' tumor suppressor protein (WT1)) disrupt the conformation of the bound linker, increase its flexibility, and impair DNA binding, thereby altering both the biological function and subnuclear localization of the protein (*76*). This model illuminates the sophisticated relationship between the function of a TF, its domain structure, and ID.

Our analysis revealed that a significant amount of DBDs are ordered or partly ordered. It is reasonable to assume that the well-folded structures of many DBDs can provide a stable platform to contact and bind specific DNA sequences. Unlike most known DBD motifs, two specific motifs, AT hook and basic domain, are totally disordered (Table 2). It is known that these two types of DBDs act as a versatile minor groove tether to anchor TFs to particular DNA sites. In addition to having sequence specific DNA-binding activity, many TFs contain a region involved in activating the transcription of the gene whose promoters or enhancers have bound. Usually, this trans-activating region enables the TF to interact with a protein involved in binding RNA polymerase. Our results have suggested that many transcriptional activation domains are either unstructured or partly structured. This is in agreement with the accumulated experimental evidence that many trans-activating domains are significantly disordered in an unbound form and that their interactions with their targets involve coupled folding and binding events (*2*, *5*, *6*, *77*). Recently, several groups comprehensively examined the kinase-inducible activation domain of CREB (cAMP response element binding protein) (*78*), the trans-activation domain of p53 (*79*), and the acidic activation domain of herpes simplex virus VP16 (*80*). These studies revealed that the activation domains remained unstructured in their normal functional states and formed a helix or helices upon binding to the target proteins.

In addition to bulk disorder, α-MoRFs are predicted to occur at a high rate within TFs, ~2.5-fold more frequent than in RandomAC_NR25. Although we were unable to link these predictions with specific functions, a consideration of previous results suggests the role that α-MoRFs may play in TF function (*21*). In addition to transcriptional regulation activity and other functions, α-MoRFs are associated with structural components of the cytoskeleton (*21*). The common thread between the cytoskeleton and TFs is the enormous macromolecular complexes in which they function. Specifically, whether bound to the cytoskeleton or to chromatin, bound proteins have a small diffusion search space in which to recruit other binding partners. For a completely ordered protein, this search space is close to zero, but for proteins with intrinsically disordered regions, the search space can be as sizable as the length of the disordered region. Disorder-to-order transition regions also have the advantage of the decoupling of specificity and affinity, demonstrated by Petros et al. 2000 (*81*), and the potential for binding to multiple binding partners (*4*). This reasoning and the high rate of α-MoRF prediction in the TF datasets suggest that α-MoRFs and MoRFs, in general, may play a general role in the function of TFs.

*Disorder and the Molecular Network.* It is known that proteins, nucleic acids, and small molecules form a dense network of molecular interactions in a cell, where molecules can be considered to be nodes and the interactions between them, edges. The architecture of a molecular network can reveal important principles of cellular organization and function, similar to how the analysis of a protein structure and ID tells us about its function and organization (*82*). The regulation of gene expression can be modeled in these complex molecular networks, with DNA-binding TFs as important components in such networks. Recently, it has been reported (*83*, *84*) that the more highly connected a protein node is (i.e., the more physically interacting partners it has), the more important it is for normal cellular function and the more likely that its removal will be lethal to a cell. Because one of the major functional advantages for ID proteins is the ability to bind to multiple, different targets without sacrificing specificity, thus forming the flexible nets (*16*), and because ID is responsible for the binding diversity of the numerous protein−protein interactions, it is reasonable to propose that disordered TFs are prime candidates for being essential protein hubs for controlling many aspects of biological activity. Our results found that 4 wholly ID proteins (HMGI-14, HMCI-C, SOX-15, and SOX-3) among the top 15 disordered TFs belong to one superfamily, the high mobility group (HMG). The HMG family could be a typical example for this model of ID proteins as interaction hubs.

It is well-known that all members of the HMGA family are characterized by the presence of three similar but independent copies of a conserved DNA-binding peptide motif (P-R-G-R-P) named AT hook. Various physical studies, including NMR analysis of a co-complex of individual AT hooks with a synthetic DNA substrate (*59*), have elucidated the physical basis for the recognition of the minor groove of AT-DNA by HMGA proteins. These studies also have demonstrated that the intrinsic flexibility of the unstructured HMGA proteins is a critical factor for substrate recognition.

In addition to their unique multiple AT hook DNA-binding characteristics, another principal reason the HMGA proteins are able to physically interact with a large number of other proteins, most of which are TFs, is that their intrinsic flexibility allows them to undergo reversible disorder-to-order structural transitions upon binding to their partners that in turn can induce conformational changes in bound DNA and protein substrates. It has been reported that at least 18 different TFs were specifically associated with the HMGA proteins as determined by various experimental methods (Figure 9). The intrinsic flexibility and binding diversity of unstructured proteins has laid down the physical foundation for HMGA to act as hubs of nuclear function and play the central role in the nucleus as sensors of a wide variety of different intra and extracellular signaling events and as integrators and effectors of the plethora of cellular responses to these stimuli (*85*). Most importantly, both the transcription of HMGA genes and the biochemical modifications of HMGA proteins are direct downstream targets of numerous signal pathways, making them exquisitely responsive to various environmental influences (*86*). Here, we have presented evidence that intrinsically disordered TFs can play crucial roles in the function of flexible transcriptional networks.

*Implications of Disorder for the Discovery of Transcription-Modulating Drugs.* In view of their pivotal roles for transcription in biological processes, TFs represent obvious targets for therapeutic drugs, which can act either by stimulating the transcription of specific genes for a desired beneficial effect or by inhibiting the transcription of genes involved in an undesirable event (*87*). Indeed, of the 50 FDA-approved best selling drugs, more than 10% target transcription, and these include such well-known drugs as salicylate and tamoxifen (*88*). The existence of such drugs validates that transcription does represent a suitable target for therapeutic drugs. Additionally, recent advances in the design, selection, and engineering of DNA-binding-proteins have led to the emerging field of designer TFs. Modular DNA-binding-protein domains, particularly zinc-finger domains, can be assembled to recognize a given sequence of DNA in a regulatory region of a targeted area. The potential of this technology to alter the transcription of specific genes, to discover new genes, and to induce phenotypes in cells and organisms is now being widely applied in the areas of gene therapy, pharmacology, and biotechnology (*89*). Without a doubt, the current structure−function drug design strategy is far from being effective and efficient. Thus, any new approach to identify and validate the drug target at an early stage will significantly accelerate drug discovery and development. The information on ID in TFs would be extremely useful for target selection, small molecule design, assay development, and so forth. Disorder predictors can help identify local domains within longer regions of disorder that would be amenable to structure determination. Combining the prediction of ID with other techniques provides an alternative strategy for protein structural characterization and drug target identification. For example, large regions of the tumor suppressor protein p53, another hub of cellular function and one of the most highly connected nodes in the cell, are unstructured prior to interaction with specific protein partners (*90*). The selective small-molecule (Nutlin series) antagonists of MDM2, the p53 binding partner, block the p53−MDM2 interaction and activate the p53 pathway in cancer cells, leading to cell cycle arrest, apoptosis, and growth inhibition of human tumor xenografts in nude mice (*91*).

## ACKNOWLEDGMENT

We are grateful to Drs. John Calley, Tao Wei, and Yong Yue of the Bioinformatics Group, Eli Lilly and Company for critical suggestions in statistical analysis and dataset and manuscript preparation and to Kanf Peng of Temple University for providing the software for the VSL1 disorder predictor. Thanks also go to Mat Allen, Dick Repasky, and Peng Wang from the University Information Technology Services of Indiana University for assistance with the AVIDD cluster. Finally, we thank Molecular Kinetics, Inc. for providing the software for PONDR VL-XT, α-MoRF, CDF, and CH-plot predictions.

## REFERENCES

1. Dunker, A. K., Obradovic, Z., Romero, P., Garner, E. C., and Brown, C. J. (2000) Intrinsic protein disorder in complete genomes, *Genome Inf. Ser. 11*, 161−171.
2. Ward, J. J., Sodhi, J. S., McGuffin, L. J., Buxton, B. F., and Jones, D. T. (2004) Prediction and functional analysis of native disorder in proteins from the three kingdoms of life, *J. Mol. Biol. 337*, 635−45.
3. Oldfield, C. J., Cheng, Y., Cortese, M. S., Brown, C. J., Uversky, V. N., and Dunker, A. K. (2005) Comparing and combining predictors of mostly disordered proteins, *Biochemistry 44*, 1989−2000.
4. Dunker, A. K., Lawson, J. D., Brown, C. J., Williams, R. M., Romero, P., Oh, J. S., Oldfield, C. J., Campen, A. M., Ratliff, C. M., Hipps, K. W., Ausio, J., Nissen, M. S., Reeves, R., Kang, C., Kissinger, C. R., Bailey, R. W., Griswold, M. D., Chiu, W., Garner, E. C., and Obradovic, Z. (2001) Intrinsically disordered protein, *J. Mol. Graphics Modell. 19*, 26−59.
5. Iakoucheva, L. M., Brown, C. J., Lawson, J. D., Obradovic, Z., and Dunker, A. K. (2002) Intrinsic disorder in cell-signaling and cancer-associated proteins, *J. Mol. Biol. 323*, 573−84.
6. Dyson, H. J., and Wright, P. E. (2002) Coupling of folding and binding for unstructured proteins, *Curr. Opin. Struct. Biol. 12*, 54−60.
7. Uversky, V. N., Oldfield, C. J., and Dunker, A. K. (2005) Showing your ID: intrinsic disorder as an ID for recognition, regulation and cell signaling, *J. Mol. Recognit. 18*, 343−384.
8. Wright, P. E., and Dyson, H. J. (1999) Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm, *J. Mol. Biol. 293*, 321−331.
9. Dyson, H. J., and Wright, P. E. (2005) Intrinsically unstructured proteins and their functions, *Nat. Rev. Mol. Cell Biol. 6*, 197−208.
10. Meador, W. E., Means, A. R., and Quiocho, F. A. (1992) Target enzyme recognition by calmodulin: 2.4 Å structure of a calmodulin-peptide complex, *Science 257*, 1251−1255.
11. Gunasekaran, K., Tsai, C. J., Kumar, S., Zanuy, D., and Nussinov, R. (2003) Extended disordered proteins: targeting function with less scaffold, *Trends Biochem. Sci. 28*, 81−85.
12. Fink, A. L. (2005) Natively unfolded proteins, *Curr. Opin. Struct. Biol. 15*, 35−41.
13. Dunker, A. K., and Obradovic, Z. (2001) The protein trinity−linking function and disorder, *Nat. Biotechnol. 19*, 805−806.
14. Tompa, P. (2002) Intrinsically unstructured proteins, *Trends Biochem. Sci. 27*, 527−533.
15. Daughdrill, G. W., Pielak, G. J., Uversky, V. N., Cortese, M. S., and Dunker, A. K. (2005) Natively disordered proteins, in *Handbook of Protein Folding* (Buchner, J., and Kiefhaber, T., Eds.) pp 271−353, Wiley-VCH, Verlag GmbH & Co. Weinheim, Germany.
16. Dunker, A. K., Cortese, M. S., Romero, P., Iakoucheva, L. M., and Uversky, V. N. (2005) Flexible nets. The roles of intrinsic disorder in protein interaction networks, *FEBS J. 272*, 5129−5148.
17. Spolar, R. S., and Record, M. T., Jr. (1994) Coupling of local folding to site-specific binding of proteins to DNA, *Science 263*, 777−784.
18. Shoemaker, B. A., Portman, J. J., and Wolynes, P. G. (2000) Speeding molecular recognition by using the folding funnel: the fly-casting mechanism, *Proc. Natl. Acad. Sci. U.S.A. 97*, 8868−8873.
19. Williamson, J. R. (2001) Proteins that bind RNA and the labs who love them, *Nat. Struct. Biol. 8*, 390−391.
20. Kalodimos, C. G., Biris, N., Bonvin, A. M., Levandoski, M. M., Guennuegues, M., Boelens, R., and Kaptein, R. (2004) Structure and flexibility adaptation in nonspecific and specific protein-DNA complexes, *Science 305*, 386−389.
21. Oldfield, C. J., Cheng, Y., Cortese, M. S., Romero, P., Uversky, V. N., and Dunker, A. K. (2005) Coupled folding and binding with alpha-helix-forming molecular recognition elements, *Biochemistry 44*, 12454−12470.
22. Garner, E., Romero, P., Dunker, A. K., Brown, C., and Obradovic, Z. (1999) Predicting binding regions within disordered proteins, *Genome Inf. Ser. 10*, 41−50.
23. Romero, P., Obradovic, Z., Li, X., Garner, E. C., Brown, C. J., and Dunker, A. K. (2001) Sequence complexity of disordered protein, *Proteins 42*, 38−48.
24. Obradovic, Z., Peng, K., Vucetic, S., Radivojac, P., and Dunker, A. K. (2005) Exploiting heterogeneous sequence properties improves prediction of protein disorder, *Proteins 61*, 176−182.
25. Uversky, V. N., Gillespie, J. R., and Fink, A. L. (2000) Why are "natively unfolded" proteins unstructured under physiologic conditions? *Proteins 41*, 415−27.
26. Hobohm, U., Scharf, M., Schneider, R., and Sander, C. (1992) Selection of representative protein data sets, *Protein Sci. 1*, 409−417.
27. Li, W., Jaroszewski, L., and Godzik, A. (2001) Clustering of highly homologous sequences to reduce the size of large protein databases, *Bioinformatics 17*, 282−283.

28. Li, W., Jaroszewski, L., and Godzik, A. (2002) Sequence clustering strategies improve remote homology recognitions while reducing search times, *Protein Eng. 15*, 643−649.

29. Li, X., Romero, P., Rani, M., Dunker, A. K., and Obradovic, Z. (1999) Predicting protein disorder for N-, C-, and internal regions. *Genome Inf. Ser. 0*, 30−40.

30. Romero, P., Obradovic, Z., and Dunker, A. K. (1997) Sequence data analysis for long disordered regions prediction in the calcineurin family, *Genome Inf. 8*, 110−124.

31. Romero, P., Obradovic, Z., Kissinger, C., Villafranca, J. E., and Dunker, A. K. (1997) Identifying disordered regions in proteins from amino acid sequence, *Proc. Int. Conf. Neural Networks 1*, 90−95.

32. Peng, K., Vucetic, S., Radivojac, P., Brown, C. J., Dunker, A. K., and Obradovic, Z. (2005) Optimizing long intrinsic disorder predictors with protein evolutionary information, *J. Bioinform. Comput. Biol. 3*, 35−60.

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

34. Rost, B., Sander, C., and Schneider, R. (1994) PHD- -an automatic mail server for protein secondary structure prediction, *Comput. Appl. Biosci. 10*, 53−60.

35. Jones, D. T. (1999) Protein secondary structure prediction based on position-specific scoring matrices, *J. Mol. Biol. 292*, 195−202.

36. Sprent, P. (1993) *Applied Nonparametric Statistical Methods*, 2nd ed., Chapman and Hall, London.

37. Callaghan, A. J., Aurikko, J. P., Ilag, L. L., Gunter Grossmann, J., Chandran, V., Kuhnel, K., Poljak, L., Carpousis, A. J., Robinson, C. V., Symmons, M. F., and Luisi, B. F. (2004) Studies of the RNA degradosome-organizing domain of the *Escherichia coli* ribonuclease RNase E, *J. Mol. Biol. 340*, 965−979.

38. Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., and Bourne, P. E. (2000) The protein data bank, *Nucleic Acids Res. 28*, 235−242.

39. Boeckmann, B., Bairoch, A., Apweiler, R., Blatter, M. C., Estreicher, A., Gasteiger, E., Martin, M. J., Michoud, K., O'Donovan, C., Phan, I., Pilbout, S., and Schneider, M. (2003) The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003, *Nucleic Acids Res. 31*, 365−370.

40. Peng, K., Obradovic, Z., and Vucetic, S. (2004) Exploring bias in the Protein Data Bank using contrast classifiers. *Pac Symp Biocomput*, 435−46.

41. Dunker, A. K., Garner, E., Guilliot, S., Romero, P., Albrecht, K., Hart, J., Obradovic, Z., Kissinger, C., and Villafranca, J. E. (1998) Protein disorder and the evolution of molecular recognition: theory, predictions and observations, *Pac. Symp. Biocomput. '98*, 473−484.

42. Vucetic, S., Brown, C. J., Dunker, A. K., and Obradovic, Z. (2003) Flavors of protein disorder, *Proteins 52*, 573−584.

43. Vihinen, M. (1987) Relationship of protein flexibility to thermostability, *Protein Eng. 1*, 477−480.

44. Tupler, R., Perini, G., and Green, M. R. (2001) Expressing the human genome, *Nature 409*, 832−833.

45. Wolfe, S. A., Greisman, H. A., Ramm, E. I., and Pabo, C. O. (1999) Analysis of zinc fingers optimized via phage display: evaluating the utility of a recognition code, *J. Mol. Biol. 285*, 1917−1934.

46. Tjian, R., and Maniatis, T. (1994) Transcriptional activation: a complex puzzle with few easy pieces, *Cell 77*, 5−8.

47. Zupicich, J., Brenner, S. E., and Skarnes, W. C. (2001) Computational prediction of membrane-tethered transcription factors, *GenomeBiology 2*, RESEARCH0050.

48. Stegmaier, P., Kel, A. E., and Wingender, E. (2004) Systematic DNA-binding domain classification of transcription factors, *Genome Inf. Ser. 15*, 276−286.

49. Patel, L., Abate, C., and Curran, T. (1990) Altered protein conformation on DNA binding by Fos and Jun, *Nature 347*, 572−575.

50. Shuman, J. D., Vinson, C. R., and McKnight, S. L. (1990) Evidence of changes in protease sensitivity and subunit exchange rate on DNA binding by C/EBP, *Science 249*, 771−774.

51. O'Neil, K. T., Hoess, R. H., and DeGrado, W. F. (1990) Design of DNA-binding peptides based on the leucine zipper motif, *Science 249*, 774−778.

52. Weiss, M. A., Ellenberger, T., Wobbe, C. R., Lee, J. P., Harrison, S. C., and Struhl, K. (1990) Folding transition in the DNA-binding domain of GCN4 on specific binding to DNA, *Nature 347*, 575−578.

53. Anthony-Cahill, S. J., Benfield, P. A., Fairman, R., Wasserman, Z. R., Brenner, S. L., Stafford, W. F., III, Altenbach, C., Hubbell, W. L., and DeGrado, W. F. (1992) Molecular characterization of helix-loop-helix peptides, *Science 255*, 979−983.

54. Ferre-D'Amare, A. R., Pognonec, P., Roeder, R. G., and Burley, S. K. (1994) Structure and function of the b/HLH/Z domain of USF, *EMBO J. 13*, 180−189.

55. Reeves, R., and Nissen, M. S. (1990) The A.T-DNA-binding domain of mammalian high mobility group I chromosomal proteins. A novel peptide motif for recognizing DNA structure, *J. Biol. Chem. 265*, 8573−8582.

56. Reeves, R., and Beckerbauer, L. (2001) HMGI/Y proteins: flexible regulators of transcription and chromatin structure, *Biochim. Biophys. Acta 1519*, 13−29.

57. Reeves, R. (2004) HMGA proteins: isolation, biochemical modifications, and nucleosome interactions, *Methods Enzymol. 375*, 297−322.

58. Aravind, L., and Landsman, D. (1998) AT-hook motifs identified in a wide variety of DNA-binding proteins, *Nucleic Acids Res. 26*, 4413-21.

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

60. Brennan, R. G., and Matthews, B. W. (1989) The helix-turn-helix DNA binding motif, *J. Biol. Chem. 264*, 1903−1906.

61. Miller, J., McLachlan, A. D., and Klug, A. (1985) Repetitive zinc-binding domains in the protein transcription factor IIIA from Xenopus oocytes. *EMBO J. 4*, 1609−1614.

62. Lee, M. S., Gippert, G. P., Soman, K. V., Case, D. A., and Wright, P. E. (1989) Three-dimensional solution structure of a single zinc finger DNA-binding domain, *Science 245*, 635−637.

63. Zhou, H. X. (2001) The affinity-enhancing roles of flexible linkers in two-domain DNA-binding proteins, *Biochemistry 40*, 15069−15073.

64. Wuttke, D. S., Foster, M. P., Case, D. A., Gottesfeld, J. M., and Wright, P. E. (1997) Solution structure of the first three zinc fingers of TFIIIA bound to the cognate DNA sequence: determinants of affinity and sequence specificity, *J. Mol. Biol. 273*, 183−206.

65. Foster, M. P., Wuttke, D. S., Radhakrishnan, I., Case, D. A., Gottesfeld, J. M., and Wright, P. E. (1997) Domain packing and dynamics in the DNA complex of the N-terminal zinc fingers of TFIIIA, *Nat. Struct. Biol. 4*, 605−608.

66. Bowers, P. M., Schaufler, L. E., and Klevit, R. E. (1999) A folding transition and novel zinc finger accessory domain in the transcription factor ADR1, *Nat. Struct. Biol. 6*, 478−485.

67. van Leeuwen, H. C., Strating, M. J., Rensen, M., de Laat, W., and van der Vliet, P. C. (1997) Linker length and composition influence the flexibility of Oct-1 DNA binding, *EMBO J. 16*, 2043−2053.

68. Peisach, E., and Pabo, C. O. (2003) Constraints for zinc finger linker design as inferred from X-ray crystal structure of tandem Zif268-DNA complexes, *J. Mol. Biol. 330*, 1−7.

69. Jantz, D., and Berg, J. M. (2004) Reduction in DNA-binding affinity of Cys2His2 zinc finger proteins by linker phosphorylation, *Proc. Natl. Acad. Sci. U.S.A. 101*, 7589−7593.

70. Bustin, M. (2001) Revised nomenclature for high mobility group (HMG) chromosomal proteins, *Trends Biochem. Sci. 26*, 152−153.

71. Gerstein, M. (1998) How representative are the known structures of the proteins in a complete genome? A comprehensive structural census. *Folding Des. 3*, 497−512.

72. Brenner, S. E. (2000) Target selection for structural genomics, *Nat. Struct. Biol. 7*, 967−969.

73. Park, J., Holm, L., Heger, A., and Chothia, C. (2000) RSDB: representative protein sequence databases have high information content, *Bioinformatics 16*, 458−464.

74. Patikoglou, G., and Burley, S. K. (1997) Eukaryotic transcription factor-DNA complexes, *Annu. Rev. Biophys. Biomol. Struct. 26*, 289−325.

75. Laity, J. H., Dyson, H. J., and Wright, P. E. (2000) DNA-induced alpha-helix capping in conserved linker sequences is a determinant of binding affinity in Cys(2)-His(2) zinc fingers, *J. Mol. Biol. 295*, 719−27.
76. Laity, J. H., Dyson, H. J., and Wright, P. E. (2000) Molecular basis for modulation of biological function by alternate splicing of the Wilms' tumor suppressor protein, *Proc. Natl. Acad. Sci. U.S.A. 97*, 11932−11935.
77. Frankel, A. D., and Kim, P. S. (1991) Modular structure of transcription factors: implications for gene regulation, *Cell 65*, 717−719.
78. Radhakrishnan, I., Perez-Alvarado, G. C., Parker, D., Dyson, H. J., Montminy, M. R., and Wright, P. E. (1997) Solution structure of the KIX domain of CBP bound to the transactivation domain of CREB: a model for activator:coactivator interactions, *Cell 91*, 741−752.
79. Ayed, A., Mulder, F. A., Yi, G. S., Lu, Y., Kay, L. E., and Arrowsmith, C. H. (2001) Latent and active p53 are identical in conformation, *Nat. Struct. Biol. 8*, 756−760.
80. Grossmann, J. G., Sharff, A. J., O'Hare, P., and Luisi, B. (2001) Molecular shapes of transcription factors TFIIB and VP16 in solution: implications for recognition, *Biochemistry 40*, 6267−6274.
81. Petros, A. M., Nettesheim, D. G., Wang, Y., Olejniczak, E. T., Meadows, R. P., Mack, J., Swift, K., Matayoshi, E. D., Zhang, H., Thompson, C. B., and Fesik, S. W. (2000) Rationale for Bcl-xL/Bad peptide complex formation from structure, mutagenesis, and biophysical studies, *Protein Sci. 9*, 2528−2534.
82. Spirin, V., and Mirny, L. A. (2003) Protein complexes and functional modules in molecular networks, *Proc. Natl. Acad. Sci. U.S.A. 100*, 12123−12128.
83. Jeong, H., Tombor, B., Albert, R., Oltvai, Z. N., and Barabasi, A. L. (2000) The large-scale organization of metabolic networks, *Nature 407*, 651−654.
84. Jeong, H., Mason, S. P., Barabasi, A. L., and Oltvai, Z. N. (2001) Lethality and centrality in protein networks, *Nature 411*, 41−42.
85. Reeves, R. (2001) Molecular biology of HMGA proteins: hubs of nuclear function, *Gene 277*, 63−81.
86. Reeves, R. (2000) Structure and function of the HMGI(Y) family of architectural transcription factors, *Environ. Health Perspect. 108*, 803−809.
87. Latchman, D. S. (2000) Transcription factors as potential targets for therapeutic drugs, *Curr. Pharm. Biotechnol. 1*, 57−61.
88. Cai, W., Hu, L., and Foulkes, J. G. (1996) Transcription-modulating drugs: mechanism and selectivity, *Curr. Opin. Biotechnol. 7*, 608−615.
89. Blancafort, P., Segal, D. J., and Barbas, C. F., III. (2004) Designing transcription factor architectures for drug discovery, *Mol. Pharmacol. 66*, 1361−1371.
90. Kussie, P. H., Gorina, S., Marechal, V., Elenbaas, B., Moreau, J., Levine, A. J., and Pavletich, N. P. (1996) Structure of the MDM2 oncoprotein bound to the p53 tumor suppressor transactivation domain, *Science 274*, 948−953.
91. Vassilev, L. T., Vu, B. T., Graves, B., Carvajal, D., Podlaski, F., Filipovic, Z., Kong, N., Kammlott, U., Lukacs, C., Klein, C., Fotouhi, N., and Liu, E. A. (2004) In vivo activation of the p53 pathway by small-molecule antagonists of MDM2, *Science 303*, 844−848.
92. Hobohm, U., and Sander, C. (1994) Enlarged representative set of protein structures, *Protein Sci. 3*, 522−524.

BI0602718