Intrinsic Disorder in Transcription Factors

Jiangang (Al) Liu

Submitted to the faculty of the Indiana University School of Informatics Graduate School in partial fulfillment of the requirements for the degree Master of Sciences in Bioinformatics, August 2005

TABLE OF CONTENTS

TABLE OF CONTENTS (continued)

I. ACKNOWLEDGEMENTS

Committee members:

Dr. Narayanan B Perumal

Dr. Vladimir Uversky

Dr. A Keith Dunker

Eli Lilly and Company:

Dr. Eric W Su

Dr. John Calley

Center for Computational Biology and Bioinformatics at IUPUI:

Dr. Chris Oldfield and Dr. Pedro Romero

Jie Sun, Andrew Campen, Amrita Mohan

My family:

Jeffrey, Fangzhou, and Hongling Xiao

II. ABSTRACT

Reported evidence suggested that high abundance of intrinsic disorder in eukaryotic genomes in comparison to bacteria and archaea may reflect the greater need for disorder-associated signaling and transcriptional regulation in nucleated cells. The major advantage of intrinsically disordered proteins or disordered regions is their inherent plasticity for molecular recognition, and this advantage promotes disordered proteins or disordered regions in binding their targets with high specificity and low affinity and with numerous partners. Although several well-characterized examples of intrinsically disordered proteins in transcriptional regulation have been reported and the biological functions associated with their corresponding structural properties have been examined, so far no specific systematic analysis of intrinsically disordered proteins has been reported. To test for a generalized prevalence of intrinsic disorder in transcriptional regulation, we first used the Predictor Of Natural Disorder Regions (PONDR VL-XT) to systematically analyze the intrinsic disorder in three Transcription Factor (TF) datasets (*TFSPTRENR25, TFSPNR25, TFNR25*) and two control sets (*PDBs25 and RandomACNR25*). PONDR VL-XT predicts regions of ≥30 consecutive disordered residues for 94.13%, 85.19%, 82.63%, 54.51%, and 18.64% of the proteins from *TFNR25, TFSPNR25, TFSPTRENR25*, *RandomACNR25*, and *PDBs25*, respectively, indicating significant abundance of intrinsic disorder in TFs as compared to the two control sets. We then used Cumulative Distribution Function (CDF) and charge-hydropathy plots to further confirm this propensity for intrinsic disorder in TFs. The amino acid compositions results showed that the three TF datasets differed significantly

from the two control sets. All three TF datasets were substantially depleted in *order-promoting* residues such as ***W, F, I, Y***, and ***V***, and significantly enriched in *disorder-promoting* residues such as ***Q, S***, and ***P***. ***H*** and ***C*** were highly over-represented in TF datasets because nearly a half of TFs contain several zinc-fingers and the most popular type of zinc-finger is ***C2H2***. High occurrence of proline and glutamine in these TF datasets suggests that these residues might contribute to conformational flexibility needed during the process of binding by co-activators or repressors during transcriptional activation or repression. The data for disorder predictions on TF domains showed that the AT-hooks and basic regions of DNA Binding Domains (DBDs) were highly disordered (the overall disorder scores are 99% and 96% respectively). The C2H2 zinc-fingers were predicted to be highly ordered; however, the longer the zinc finger linkers, the higher the predicted magnitude of disorder. Overall, the degree of disorder in TF activation regions was much higher than that in DBDs. Our studies also confirmed that the degree of disorder was significantly higher in eukaryotic TFs than in prokaryotic TFs, and the results reflected the fact that the eukaryotes have well-developed elaborated gene transcription mechanism, and such a system is in great need of TF flexibility. Taken together, our data suggests that intrinsically disordered TFs or partially unstructured regions in TFs play key roles in transcriptional regulation, where folding coupled to binding is a common mechanism.

III. INTRODUCTION

III.A TRANSCRIPTION FACTORS

Regulation of gene expression is essential to the normal development and proper functioning of complex organisms. Such regulation is primarily achieved at the level of gene transcription wherein the DNA is copied into an RNA transcript. To accomplish this process in eukaryotic cells, it requires three different RNA polymerases (*RNA Pol*). Each RNA Pol is responsible for a different class of transcription: *Pol I* transcribes rRNA (ribosome RNA), *Pol II* translates for mRNA (messager RNA), and *Pol III* is for tRNA (transfer RNA) and other small RNAs. Although the control of gene regulations occurs in multiple steps, the overwhelming majority of regulatory events occur at transcription initiated by *Pol II*. However, *Pol II* cannot initiate transcription in eukaryotic genes on their own, and they absolutely require additional proteins to be involved. Any protein that is needed for the initiation of transcription is defined as a transcription factor (TF).

In general, transcription factors (TFs) are divided into two groups [Villard J. 2004]. (1) *Basal transcription factors*: they are ubiquitous and required for the initiation of RNA synthesis at all promoters. With RNA PolII, they form a complex called the basal transcription apparatus surrounding the transcription start point, and they determine the site of initiation. (2) *Gene-specific transcription factors*: this is a group of proteins that activate or repress basal transcription. These proteins are able to bind to specific DNA sequences (transcription factor binding sites, TFBS) in the gene promoter and upstream of the transcription start site (TSS), and act in concert with co-activator or co-repressor proteins to activate or inhibit transcription. These proteins bind to regulatory sequences organized in a series of regulatory modules along the DNA. Thus, the molecular basis for

transcriptional regulation of gene expression is the binding of trans-acting proteins (transcription factors) to *cis*-acting sequences (binding sites) [Villard J. 2004].

The critical importance of gene regulation by transcription is indicated by the fact that a growing list of human diseases is due to genetic defects in TFs. It ranges from developmental syndromes such as the mutation of POU4F3 that causes hearing loss [de Kok YJ, *et al*; 1995] to a wide variety of more common sporadic human cancers like various carcinomas, brain tumors, sarcomas, and leukemia [Sherr CJ and McCormick F, 2002].

In view of its pivotal role for transcription in biological processes, TF represents an obvious target for therapeutic drugs which could 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 [Latchman D.S, 2000]. 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 [Cai W. *et al*, 1996]. The existence of such drugs indicates 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 phenotype in cells and organisms is now being widely applied in the areas of gene therapy, pharmacology, and biotechnology [Blancafort *et al*, 2004].

The last two decades have witnessed a tremendous expansion in our knowledge of TFs and their roles employed by eukaryotic cells to control gene activity. Ample evidence has accumulated that show eukaryotic TFs contain a variety of structural motifs that interact with specific DNA sequences. Besides the *cis* elements, some promoter elements, such as TATA, GC, and CCAAT boxes, are common sequence elements to control transcription of many genes. In addition to having such as a sequence-specific DNA-binding motif, TF contains a region involved in activating the transcription of the gene whose promoters or enhancers they have bound. Usually, this trans-activating region enables the TF to interact with a protein involved in binding RNA polymerase (*Figure 1*). Based on the comparison between the sequences of many available transcription factors and deletion

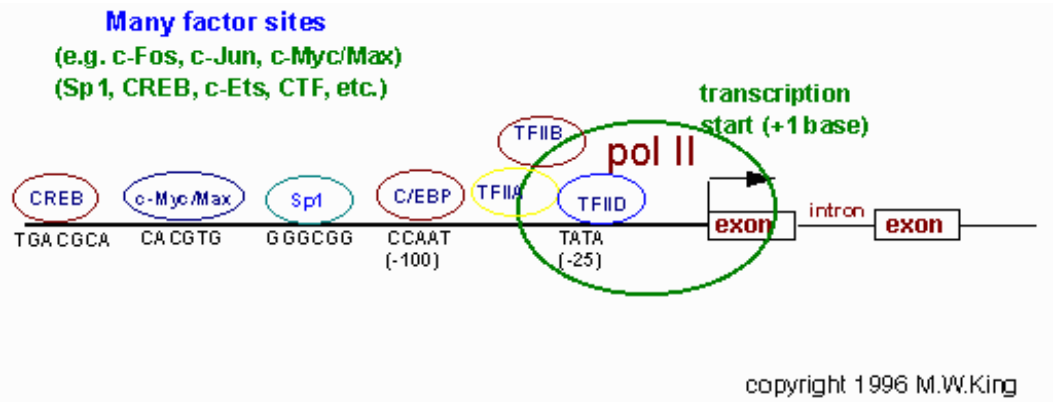


Figure 1: The diagram indicates the TATA-box and CCAAT-box basal elements at positions -25 and -100, respectively. The transcription factor TFIID has been shown to be the TATA-box binding protein, TBP. Several additional transcription factor binding sites have been included and shown to reside upstream of the 2 basal elements and of the transcriptional start site. The large green circle represents RNA polymerase II.

(mutation) analysis, several common types of motifs or functional regions in TFs have been found. These are

- DNA-binding domain (DBD)
- Trans-activation or activation domain
- Linker domain
- Dimerization domain
- Nuclear localization domain
- Ligand binding domain

It is known that not all TFs bind DNA – some just bind other transcription factors; not all TFs activate transcription – some indeed repress it; not all TFs have all functional domains listed above; characteristically, most TFs share two common functional components: a DNA-binding domain and a trans-activation domain.

III.A.1 DNA Binding Domains

Extensive structure studies of the isolated subdomains and the intact domain, employing both NMR spectroscopy and X-ray diffraction, have established the folding topology of the DNA –Binding Domain (DBD). As far as we know, most TFs share a common framework structure in their respective DNA binding sites, and slight differences in the amino acids at the binding site can alter the sequence of the DNA to which it binds. Several attempts for classifying TFs on the basis of their DNA-binding were published recently and none has the perfect solution [Stegmaier P. et al 2004]. For example, many TF members were found which could not yet be assigned to any group at all, and some members require re-assignment due to either discovering a new insight of structural features of many DBDs or increasing knowledge about the complexity of TF domain composition. Based on the current knowledge about the structural difference of

TF-DNA complexes, TFs are generally classified into several families, and those here are just some of the main types [Patikoglou G and Burley, SK, 1997; Stegmaier P. et al 2004]

Helix-Turn-Helix: This group of proteins is well known to contain the helix-turn-helix motif. It includes: *(1) Homeodomain Proteins (Figure 2):* this is a set of very important family of TFs. The homeodomain consists of 60 amino acids arranged in a helix-turn-helix. Its third helix extends to the major groove of the DNA that it recognizes.

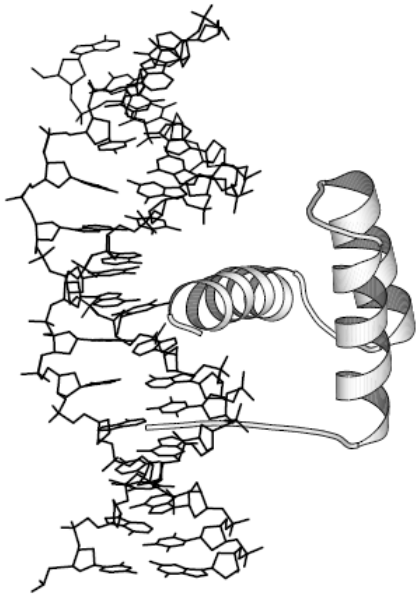


Figure 2.The homeodomain is a compact 60-residue DBD that consists of three α-helices folded around hydrophobic core and a flexible N-terminal arm that becomes ordered only on DNA binding (Kissinger et al., 1990; Patikoglou & Burley 1997)

The amino-terminal portion of the homeodomain is a flexible arm that becomes ordered only when its arm binds on the DNA at the base of the minor groove [Kissinger CR, et al, 1990]; *(2) Myb:* this family of TFs consists of three imperfect direct repeats, the second and third of which are responsible for sequence specific DNA binding. The third repeat of *c-Myb* includes three α-helices folded around a tryptophan-rich hydrophobic core and

resembles the HTH homeodomain [Ogata K, et al 1992]*; (3) Winged-Helix/fork head Proteins*: this is the group of TFs that contains a highly conserved 110-residue DNA-binding domain. The winged-helix motif binds DNA by presenting the recognition helix to the major groove, with two wing-like loops interacting with flanking portions of the phosphoribose backbone and the adjacent minor groove [Clark KL, et al 1993]; *(4) Cap-like domains*: these domains are the DNA-binding portions of various eukaryotic transcriptional activators, including heat shock factor and ETS-containing proteins [Steitz TA. 1990]. *(5) POU domains*: These proteins contain two discrete recognition helices to the major groove, including a POU homeodomain and a POU-specific domain, that are tethered to one another by a flexible linker that is not visible in the electron density map [Klemm JD, et al 1994]. *(6) Paired domains*: These domains are found in Pax proteins. It contains two globular sub-domains that both resemble the homeodomain. Unlike the POU domain, only the N-terminal subdomain presents it recognition helix to the major groove.

Basic domains: This class of TFs possesses a specific domain characterized by a large excess of positive charges, preventing them from being structured when free in solution, but becoming α-helically folded when interacting with DNA [Weiss, M.A. et al, 1990]. It functions as a homo- and /or heterodimer, using its α-helical basic regions to grip the double helix in a scissors-like fashion, making sequence-specific side chain-base contacts in the major groove. For example, *(1) Basic region /Leucine zipper proteins*: contains a characteristic leucine zipper potion to form a canonical left-handed, α-helical coiled-coil. X-ray and NMR solution confirmed the scissors grip model. Circular dichroism spectroscopy observed that the basic region undergoes a random coil-to-a-helix

folding transition when it binds to its cognate DNA [Saudek V, et al 1991]. *(2) Basic region/helix-loop-helix/leucine zipper proteins*: These are characterized by a highly conserved 60-100 residues motif comprised of two amphipathic α-helices separated by a loop of variable length. The helix-loop-helix motif is primarily responsible for dimerization. Most of these proteins posses a highly conserved basic region, that mediates high-affinity, sequence-specific DNA binding. The basic region undergoes a random coil-to-a-helix folding transition via an induced-fit mechanism to bind its cognate DNA.

Zinc-Coordinating domains: Many eukaryotic DNA-binding proteins contain zinc as an essential cofactor. Zinc-binding proteins account for nearly half of TFs in the human genome and are the most abundant class of proteins in human proteome [Tuple et al, 2001] (*Figure 3*). They could be roughly classified into following five classes: *(1)*

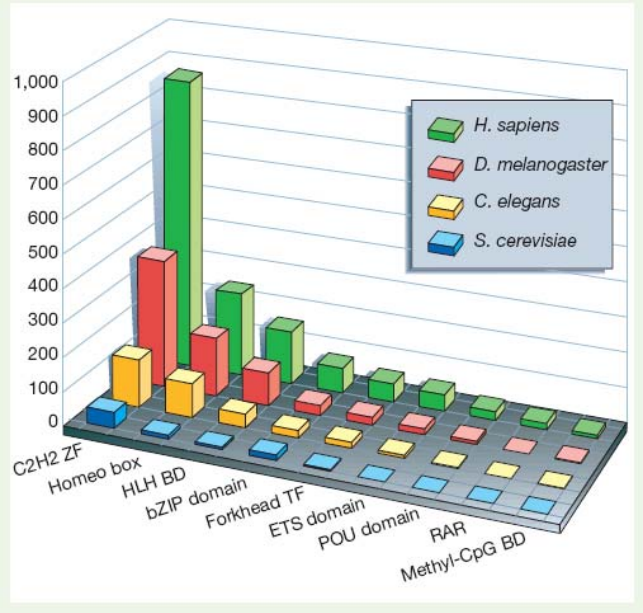


Figure 3: The TF families shown are the largest of their category out of the 1502 human Protein families listed by the IPI (Tupler et al., Nature, 409: 832-833)

Transcription factor IIIA: It was identified as a repeated 27-residue motif that contains a zinc-binding Cys2His2 tetrad, now referred to as a zinc-finger. The Zif268-DNA complex structure revealed three zinc fingers, each presenting their recognition helices to the major groove (*Figure 4a*); *(2) Steriod / nuclear receptor proteins*: They are

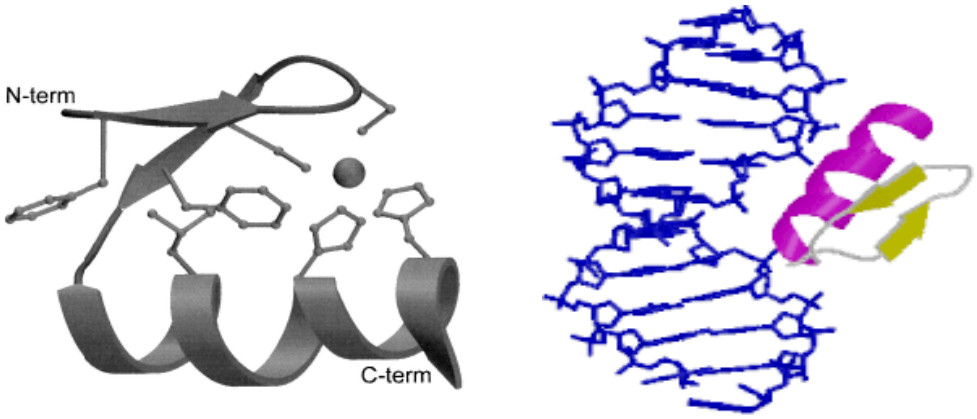


Figure 4a: Diagram of the motif from finger 2 of Zif268 (37, 95). The side chains of the conserved cysteines and histidines, which are involved in zinc coordination, and side chains of the three conserved hydrophobic residues are shown.
and by the conserved hydrophobic core that flanks the zinc binding site (84, 113).

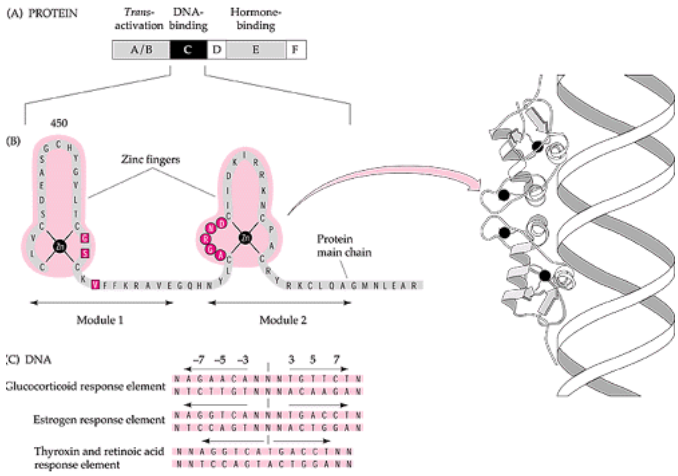


Figure 4b: Structural organization of hormone receptor DNA-binding proteins
(**A**) Generalized structure of a steroid hormone binding protein. receptor proteins.
(**B**) Zinc finger DNA-binding region of the glucocorticoid receptor.
(**C**) The zinc finger region of the glucocorticoid receptor bound to its responsive element.
(After Kaptein, 1992.) (*http://www.devbio.com/printer.php?ch=5&id=39*)

characterized by the presence of a Cys4 double loop-zinc helix DNA-binding motif (*Figure 4b*); *(3) GAL4 and PPR1*: A large family of Zinc-binding Cys6 TFs found in fungi; ***(4)*** *GATA*-1: This is an erythroid-specific transcription factor where DNA-binding domain is a small Cys4 Zn++ containing α /β motif, and has a α-helix in the major groove of the recognition element. In addition, random coil regions participate in interactions with both the major and minor grooves; *(5) p53*: An α-helix and a loop are presented to the major groove and make side chain-base interaction, and an arginine side chain from another loop projects into the minor groove, making phosphoribose contacts.

β-Scaffold Domains with Minor Groove Contacts: It is hard to find a common DBD in this group of TFs. Any pair or subgroup may share some characteristic, but not all of them share the common feature, "β-Scaffold**".** Several TFs that contain the β-scaffold feature are listed here: *(1) E2:* It has a dimeric eight-stranded antiparallel β-barrel structure; *(2)REL:* Differing from other TFs, the complete 300-amino acid *Rel* homology region is required for DNA recognition, and residues from the entire length of the protein contribute to the DNA-interaction surface; *(3) Serum response factor:* contains a conserved DNA-binding region-a MADS box that forms a α-β-α sandwich, which presents an antiparallel coiled –coil to the major grooves flanking a compressed minor groove in the center of a smoothly bent DNA element. Minor groove contacts are supported by the two N-termini that extend away from the helices bound in the major groove. *(4) DNA-bending proteins:* In addition to DNA-binding proteins, there is also a set of TFs that function primarily as DNA-bending proteins. Most of these proteins are characterized by a DNA-binding element called the HMG box, a set of approximately 80 amino acids that mediate the binding of these proteins to the minor groove of the DNA.

These proteins include the Y chromosome sex-determining factor, SRY, the lymphocyte enhancer protein LEF-1, and the chromatin proteins HMG-1 (Y) and HMG-2. These proteins are not thought to activate transcription by directly interacting with the transcription apparatus. Rather, they are thought to bend the DNA so that the activators and repressors can be brought into contact [Reeves R and Beckerbauer L. 2001].

III.A.2 TF activation domain

Compared to the DNA-binding domain, much less is known about another functional component of TFs: activation domain. Current evidence indicates that following DNA binding, a transcription factor exerts an influence over gene expression mediated through the trans-activation domain. In most cases activation domain ranges from 30-100 amino acids in length and contains variable functional amino acid arrangements such as acid block sequence with high concentration of negatively charged residues [Ptashne, 1988], glutamine- or proline-rich regions [Mitchell and Tijan, 1989]. Activation domains may act directly, or they may recruit a co-activator that possesses activation properties and an ability to interact with the basal transcription complex, but lacks any intrinsic DNA-binding capacity.

III.B INTRINSIC DISORDER AND PROTEIN FUNCTION

Interpretation of protein function in terms of specific three-dimensional structure has dominated the thinking about proteins for more than 100 years. This concept, as a lock-and key model for elucidating the specificity of the enzymatic hydrolysis of glucosides [Fischer 1894], proved to be extremely fruitful. However, the {sequence} -> {3 D Structure} -> {Function} paradigm is simply not true for many proteins. Numerous counterexamples have surfaced over the years – proteins which lack three –dimensional

structure still can perform biological functions [Dunker and Obradovic 2001]. Over the past decade, Dunker and several others have made a pioneering discovery that, under physiological conditions (pH 7.0 and 25˚C), some proteins and protein domains exist with little or no ordered structure [Dunker *et al.* 2002; Uversky, V. *et al*. 2000]. These proteins have often been referred to as 'natively denatured/unfolded' or 'intrinsically unstructured /disordered' [Gunasekaran K. et al 2003]. These disordered proteins lack a folded structure but display a highly flexible, random-coil-like conformation under physiological conditions. The cumulated experimental data shows that the intrinsic disorder proteins do not possess uniform structural properties, rather these proteins can exist in any one of the three thermodynamic states in term of the Protein Trinity Model: ordered forms, molten globules, and random coils [Dunker *et al*., 2001] (*Figure 5*). The key point of the Protein Trinity is that a particular function might depend on any one of these states or a transition between two of the states. Based on this Trinity model, Uversky added one more state called premolten globule and named a Protein Quartet Model. Proteins in premolten globule state are essentially more compact, exhibiting some amount of residual secondary structure, although they are still less dense than native or molten globule proteins. As order, molten globule, premolten globule, and random coil conformation possess defined structural differences, they could be characterized by the following experimental approaches and applications [for recent review see Uversky *et al* 2005]:

III.B.1 Experimental Approaches

X-ray crystallography: It can define the missing electron density structures in many protein structures, which may correspond to disordered region. The absence of

Dunker's Trinity Model
(Dunker & Obradovic: nature biotech, 19, 805-806, 2001)

Uversky's Quartet Model
(Uversky et al. Protein Science, 11: 739-756)

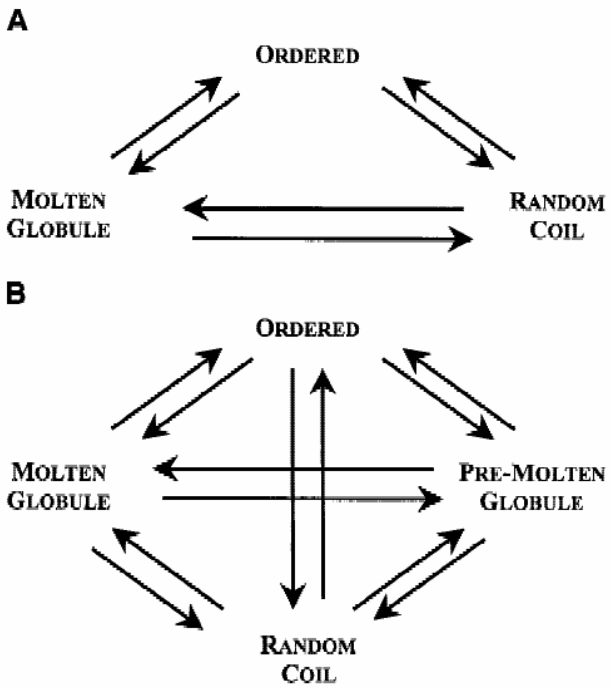


Figure 5: Two models for the continuum of protein structure

interpretable electrion density for some sections of the structure is usually associated with the increased mobility of atoms in these regions, which leads to the noncoherent X-ray scattering, making atoms invisible [Ringe & Petsko, 1986]. However, one major uncertainty regarding the information from X-ray diffraction is that, without additional experiments, it is unclear whether a region of missing electron density is a wobbly \domain, is intrinsically disordered, or just is the result of technical difficulties.

NMR spectroscopy: 3D structures can be determined for proteins in solution by NMR. Recent advances in this technology can provide detailed insights into the structure and dynamics of unfolded and partly folded states of proteins [Dyson & Wright 2002, 2005].

Circular Dichroism (CD) spectroscopy: Near – UV CD shows sharp peaks for aromatic groups when the protein is ordered, but these peaks disappear for disorder proteins due to motional averaging [Fasman GD *et al* 1996]. It, therefore, can be used to detect the intrinsic disorder protein. However, this method is only semi-quantitative and lacks residue-specific information and so does not provide clear information for proteins that contain both ordered and disordered regions.

Protease digestion: Recent studies by Fontana *et al* provide compelling evidence that flexibility, not mere surface exposure, is the major determinant for digestion at possible cut sites. Large increasing in digestion rate has been observed after the F helix of myoglobin is converted to a disordered state in apomyoglobin. Thus, hypersensitivity to proteases is clearest evidence of protein disorder. It can give position-specific information. However, the requirement for protease-sensitive residues limits the demarcation of order/disorder boundaries by this method.

Others: Several methods can be applied to indirectly detect the disorder state, including (*a*) diminished ordered secondary structure detected by several spectroscopic techniques [Smyth *et al*., 2001; Uversky *et al*., 2002], (*b*) the intermolecular mobility, solvent accessibility and compactness of a protein extracted from the analysis of different fluorescence characteristics, (*c*) the hydrodynamic parameters obtained by gel-filtration, viscometry, sedimentation, *etc*, (*d*) the information for protein conformational stability obtained by experiments, (*e*)H/D (Hydrogen/Deuterium) exchange, mass spectrometry and limited proteolysis.

Although intrinsic disorder in proteins apparently represents a common phenomenon, the number of experimentally characterized intrinsic disorder proteins is

still relatively low. One contributing factor is that the traditional biochemical methods used to produce and characterize proteins are strongly biased towards folded, ordered proteins. However, recent technology, particularly the spectroscopic methods such as NMR, has advanced in sensitivity and resolution in detecting the structural propensities and dynamics of sizeable disordered proteins or disordered regions.

III.B.2 Computational Approaches

Based on the sequence, hydrophobicity, and degree of compactness, as well as the estimated change in surface area that is exposed upon protease cleavage, a variety of computational approaches for examining disorder in native proteins have been developed [Bracken C, *et al* 2004]:

*PONDR (***P**redictor **O**f **N**atural **D**isordered **R**egions*)* [Romero P. *et al* 1997 and 2001]: This program is a variety of neural network predictors of disordered regions based on the local amino acid sequence, composition, flexibility, and other factors.

Hydropathy-Charge plot [Uversky, *et al* 2000]: This program gives a linear discriminant on the basis of the relative abundance of hydrophobic and charge residues to classify entire sequence (not regions) as ordered or disordered.

DISOPRED (Disorder Predictor) [Jones DT and Ward JJ, 2003]: Like PONDR, the algorithm behind this program also is a neural network. But the differences are that the inputs are derived from sequence profiles generated by PSI-BLAST instead of the direct protein sequence, and the output is filtered by using secondary structure predictions, so that regions confidently predicted as helix or sheet are not predicted as disordered.

GlobPlot (Predictor of Intrinsic Protein Disorder, Domain & Globularity) [Linding R, *et al* 2003]: It predicts disordered and globular regions on the basis of propensities for disorder assigned to each amino acid.

NORSp [Liu J, *et al* 2002]: It predicts 'regions of no regular secondary structure' or NORS, defined as long stretches of consecutive residues (>70) with few helix or sheet residues.

Using bioinformatics and data mining approach, Dunker and his colleagues have made fundamental discoveries showing that thousands of natively disordered proteins exist, representing a substantial fraction (around 25%) of the commonly used sequence databases [Dunker, A. K., *et al* 2000]. This observation has contributed to a reassessment of the assumption that tertiary structure is necessary for function [Wright P. E. and Dyson H. J. 1999].

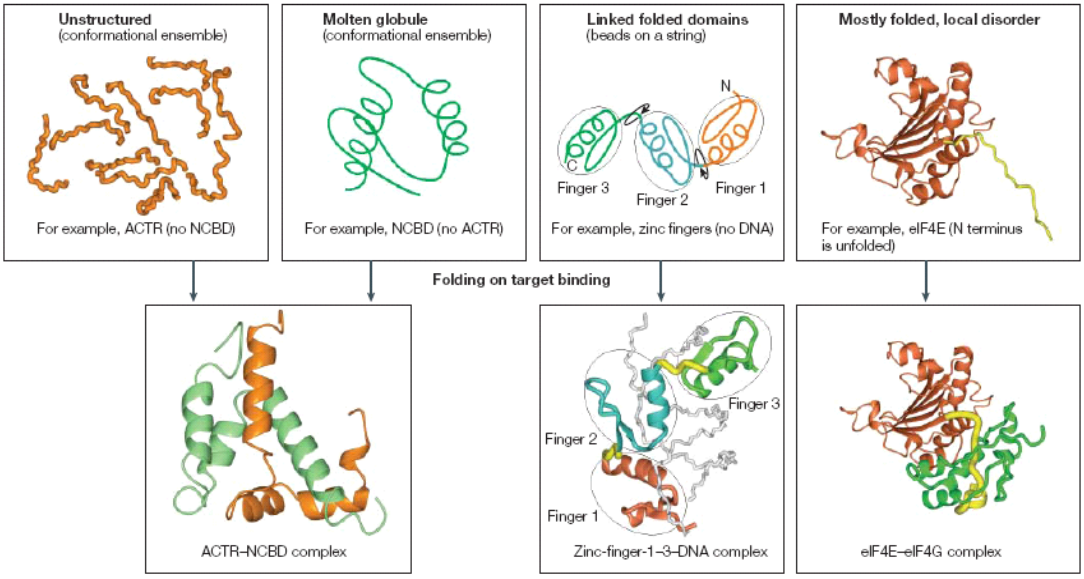
As intrinsic disorder has been better characterized, the functional importance of being disordered has been intensively analyzed. The major functional asset of disorder proteins properties, which function by molecular recognition, is related to their significant disorder-order transition. Uversky [Uversky, V. N. 2002] summarized the potential advantages of intrinsic lack of structure and function-related disorder-order transitions as described below: (1) the possibility of high specificity coupled with low affinity; (2) the ability of binding to several different targets known as one to many signaling; (3) the capability to overcome steric restrictions, enabling essentially larger interaction surfaces in the complex than could be obtained for rigid partners; (4) the precise control and simple regulation of binding thermodynamics; (5) the increased rates of specific macromolecular association; (6) the reduced lifetime of intrinsically

disordered proteins in the cell, possibly representing a mechanism of rapid turnover of the important regulatory molecules [Wright and Dyson 1999].

Although the occurrence of unstructured regions of significant size (>50 residues) is surprisingly common in functional proteins, the functional role of intrinsically disordered proteins in crucial fields such as transcriptional regulation, translation and cellular signal transduction has only recently been recognized. Since Spolar and Record first introduced the concept of binding to folding [Spolar RS and Record MT, 1994] almost a decade ago, Wright's group and others have made tremendous progress in the elucidation of the function of unstructured proteins, particularly for the crucial DNA binding and several other types of molecular recognition proteins [Dyson, H. J., and Wright, P. E. 2002; 2005]. They introduced the 'snap-lock' mechanism based on the observation of multiple zinc fingers binding to cognate DNA. Cys2His2-type zinc finger domains consist of well-folded domains connected by highly conserved linker sequences that are mobile and unstructured in the absence of the cognate DNA and behave like beads on a string. Upon binding to the correct DNA sequence, the linker becomes highly structured and locks adjacent fingers in the correct orientations in the major groove [Laity JH, *et al* 2000a]. Any alteration at this linker such as the insertion of three residues, KTS, by alternative splicing in Wilms tumor, will subsequently thwart the linker confirmation, and increases linker flexibility and impairs DNA binding, thereby both altering the biological function and sub-nuclear localization [Laity JH, *et al* 2000b].

As we have better characterized and understood the constitution of TFs in recent years, the protein–protein interactions in transcriptional regulation have become more intriguing and attractive for study. Accumulated evidence has suggested that many

transcriptional activation domains are either unstructured or partly structured, and their interactions with their targets involve coupled folding and binding events [Frankel AD and Kim PS. 1991; Dyson & Wright, 2002; Iakoucheva *et al.,* 2002; Ward *et al.,* 2004]. Recent works by several groups comprehensively examined the kinase-inducible activation domain of CREB (cAMP response element binding protein) [Radhakrishnan *et al* 1997], the trans-activation domain of p53 [Ayed A, *et al*, 2001], and the acidic activation domain of herpes simplex virus VP16 [Grossmann et al, 2001], and found that the activation domains remained unstructured in their normal functional states, and form a helix or helices upon binding to the target proteins (*Figure 6*).



Dyson and Wright: *Nature Reviews: mol. Cell Biol*. 6:197-208

Figure 6: the example of ID in TF shows the concept of coupling to folding and binding.

IV. BACKGROUND

IV.A RELATED RESEARCH

Reported evidence demonstrated that high abundance of intrinsic disorder in eukaryotic genomes in comparison to bacteria and archaea may reflect the greater need for disorder-associated signaling and regulation in nucleated cells [Dyson & Wright, 2002; Iakoucheva *et al*., 2002; Ward *et al*., 2004]. The properties of intrinsic disorder regions promote molecular recognition through the following features [Uversky 2005]: (*a*) intrinsic disorder proteins or disorder regions have the capability to combine high specificity with low affinity to couple or decouple with its functional partners; (*b*) the intrinsic plasticity enable a single disorder protein or region to recognize and bind many biological targets divisively while still being specific [Wright & Dyson, 199; Dunker *et al*., 2001; Dyson & Wright 2002, 2005]; (c) the structural propensity of intrinsic disorder proteins or disorder regions to form a large interaction faces such as the disordered region wraps-up or surrounds its partner [Meador *et al*., 1992; Dunker *et al*., 2001]; (*d*) the fast rate of association and dissociation with the important regulatory molecules to make a rapid turnover and reduce life-time of intrinsic disordered proteins in the cells. Several experimentally well-characterized proteins, such as p53, GCN4, CBP, and HMGA, interact with their partners mostly *via* regions of intrinsic disorder strongly support this concept [Dunker & Obradovic 2001].

To fully understand the abundance of intrinsic disorder in genome-wide scale of transcriptional control and to decipher the conformational structure information of TF, several attempts have been made, and the number of intrinsically disordered proteins known to be involved in cell-signaling and regulation is growing rapidly. For example,

Iakoucheva and her colleagues applied PONDR prediction on several dataset from Swiss-Prot and discovered that the intrinsic disordered proteins are prevalent in cell-signaling and cancer-associated proteins in comparison with other functional group [Iakoucheva *et al.*, 2002]. The results obtained by a different group using a predictor called DISOPRED2 to predict on *Saccharomyces* genome database, suggest that the proteins containing disorder are often located in the cell nucleus and are involved in the regulation of transcription and cell signaling [Ward *et al*., 2004]. The results also indicate that intrinsic disorder is associated with the molecular functions of kinase activity and nucleic acid binding.

It is known that most TFs have a common frame structure and share at least two functional domains: DNA-binding domain and trans-activation domain. It has been reported that the interaction of a protein with DNA often induces local folding in the protein partner [Spolar & Record, 1994]. It has been suggested that one of the important biological implications behind this coupled binding and folding scenario is that the specific signal from the complex of protein with its binding partner emerges only after appropriate conformational changes take place [Williamson, 2000].

One of the illustrative examples is the basic DNA-binding region of the leucine zipper protein GCN4 that interacts with DNA. Functional studies using circular dichroism spectroscopy document that the basic region undergoes a random coil-to-α-helix transition when it binds to its cognate AP-1 DNA site [Weiss *et al.,* 1990]. It is interesting to note that the basic region of α-helices of GCN4 was used to recognize the specific DNA binding site *via* a side chain-base contact. The DNA binding-induced folding of the basic region helix is accompanied by a considerable decrease in

conformational entropy [Bracken *et al*., 1999]. This was assumed to enhance the specificity of DNA binding, as the helical content of the basic region is greater when bound to a specific rather than to a non-specific DNA site [O'Neil *et al*., 1990].

Another well-characterized example for transcriptional activation domain is the kinase-inducible activation domain of CREB (cAMP-response element binding protein). The kinase activation domain is intrinsically disordered, both as an isolated peptide and in full-length CREB, but it folds to form a pair of helices upon binding to the KIX domain of the transcriptional co-activator CBP (CREB-binding protein) [Radhakrishnan *et al*., 1997] (also see *figure 6*).

Although several well-characterized examples of intrinsic proteins in transcriptional regulation have been reported and the biological functions associated with the corresponding structural properties have been examined [Dyson, H. J., and Wright, P. E. 2002; 2005], so far no specific systematic analysis of intrinsically disordered proteins in gene regulation has been reported.

IV.B PROSOSED HYPOTHESIS

As mentioned earlier, much higher prevalence of intrinsic disorder in eukaryotic genomes in comparison to bacteria and archaea may reflect the greater need for disorder-associated signaling and regulation in nucleated cells. The major advantage for intrinsic disorder proteins or disordered regions is their inherent plasticity for molecular recognition, and this advantage promotes disordered proteins or disordered regions in binding their targets with high specificity and low affinity and binding with numerous partners. Hence, our hypothesis proposes that since TFs are regulatory proteins interacting with DNA and multiple protein partners, they would be wholly disordered

proteins or carry regions of intrinsic disordered at a significantly higher level compared to a collection of random proteins. To prove this hypothesis in a pilot study, we randomly retrieved 10 TFs and 10 non-TF proteins respectively from Swiss-Prot for disorder prediction using PONDR VL-XT. The preliminary results looked intriguing and promising. It showed that TFs had remarkably higher average disorder score (0.56) than the random control set (0.29).

IV.C INTENDED PROJECT

In this project, we intend to apply a neural network predictor of natural disordered regions (PONDR VL-XT), cumulative distribution functions (CDFs) and charge-hydropathy plots to predict intrinsic disorder on three different TF datasets. We also aim to bring together the analysis of intrinsic disorder in TFs with a survey of the local and overall amino acid composition biases observed in TFs. A detailed computational analysis of the unstructured regions associated with its functional properties of TF domains and sub-domains will be presented. The conformational differences between DNA-binding and activation domain, and the differences of TF regions that bind DNA major groove and minor groove will be studied and analyzed. The properties of TF domain and sub-domain will be cross-examined in comparison with those observed previously using experimental approaches.

V. MATERIALS AND METHODS

V.A DATASETS

Five different data sets have been created and used for this study as described below:

V.A.1 Dataset sources and sequence retrieving methods

To construct the non-redundant, representative datasets for transcription factors from Swiss-Prot, 2683 protein sequences were downloaded from Swiss-Prot only, and total of 7195 entries were retrieved from Swiss-Prot and TrEMBL together (Swiss-Prot Release 46.2 of 01-Mar-2005, 172233 entries; TrEMBL Release 29.2 of 01-Mar-2005, 1631173 entries) by using "transcription factor" as a key word in a full-text search. *Swiss-Prot* is a curated protein sequence database that strives to provide a high level of annotation (such as the description of the function of a protein, its domain structure, post-translational modifications, variants, *etc*.) with a minimal level of redundancy and high level of integration with other databases. *TrEMBL* is a computer-annotated supplement of Swiss-Prot that contains all the translations of EMBL nucleotide sequence entries not yet integrated in *Swiss-Prot*. *Swiss-Prot/TrEMBL* in no doubt contains many more sequences than either *Swiss-Prot* or *TrEMBL* alone. However, since *TrEMBL* has not yet been integrated with *Swiss-Prot*, it may cover some sequences that already exist in *Swiss-Port*, and cause some degree of redundancy in the combining *Swiss-Prot/TrEMBL* database.

The third dataset contains1186 protein sequences, and was retrieved from the *TRANSFAC* database (TRANSFAC FACTOR TABLE, Rel.3.2 26-06-1997) (http://www.gene regulation.com/pub/databases.html#transfac) based on the availability of both sequence and domain feature. *TRANSFAC*® is a well-established database that

contains only eukaryotic cis-acting regulatory DNA elements and trans-acting factors. It covers the whole range from yeast to human. *TRANSFAC*® started in1988 with a printed compilation (Wingender, E 1988) and was transferred into computer-readable format in 1990. The TRANSFAC® data have been generally extracted from original literature; occasionally they have been taken from other compilations (Faisst and Meyer, 1992; Dhawale and Lande, 1994).

For the controls, we first obtained a dataset called *PDBs25* [Hobohm U. *et al*. 1992] from Dr. Dunker's laboratory. This set contains1771 chains with 297372 residues, and is a non-homologous subset of the structures in PDB consisting of a single representative structure for protein families whose members have < 25% sequence identity (ftp://ftp.emblheidelberg.de/pub/databases/protein_extras/pdb_select/2002_Apr.25). Although PDBs25 is a non-redundant and a well-represented (theoretically one member per family) dataset, it has been reported that trans-membrane, signal, disordered, and low complexity regions are significantly underrepresented in PDB, while disulfide bonds, metal binding sites, and sites involved in enzyme activity are overrepresented. Additionally, hydroxylation and phosphorylation, post-translational modification sites were found to be under-represented while acetylation sites were significantly overrepresented [Peng K. *et al*, 2004]. Compared to several complete genomes with a non-redundant subset of PDB, evidence indicated that the proteins encoded by the genomes were significantly different from those in the PDB with respect to sequence length, amino acid composition and predicted secondary structure composition. To overcome this bias and redundancy in PDB, we built a dataset with randomized NCBI

(GenBank) accession number. Four thousand 6-digit GenBank accession numbers were generated randomly, and then 2387 sequences were fetched from GenBank after redundant accession number elimination. We believe that this set is a representative of broad sequence diversity and reflect the natural environmental complexity.

V.A.2 Non-redundant representative dataset preparation

Biological sequence databases are highly redundant for two main reasons: (1) various databanks keep redundant sequences with many identical and nearly identical sequences; (2) natural sequences often have high sequence identities due to gene duplication from the same ancestor [Park J *et al* 2000]. It causes uneven sequence coverage and concentrates only small number of gene families and organisms. To address this problem, we have constructed four non-redundant, representative datasets called *TFNR25, TFSPNR25, TFSPTRENR25*, and *RandomACNR25*. Briefly, we first used CD-HIT (Cluster Database at High Identity with Tolerance) program from http://bioinformatics.org/project/?group_id=350 [Li *et al*., 2001, 2002] to reduce the homology to 80%, and then to 40% sequence identity according to the recommended procedures. CD-HIT is a fast and flexible program for clustering large protein databases at different sequence identity levels. This program can remove the high sequence redundancy efficiently. To achieve that no two sequences in the resulting dataset has more than 25% sequence identity, we aligned these sequences against one another using a global pair-wise sequence alignment program called *stretcher* (http://emboss.sourceforge.net/apps/stretcher.html). Two sequences with identity >25% will be stripped off from dataset. Traditionally, the sequence global alignment program using the Needleman & Wunsch algorithm, for instance, as implemented in the program

needle, requires O(MN) space and O(N) time. It will take a long time, and computer memory will rapidly be exhausted as the size of the aligned sequences increases. The *stretcher* program calculates a global alignment of two sequences using a modification of the classic dynamic programming algorithm that uses linear space. This program implements the Myers and Miller algorithm for finding an optimal global alignment in an amount of computer memory that is only proportional to the size of the smaller sequence, O(N). The computing time has been shortened from two weeks to 4 hours after the *needle* program was replaced with *stretcher* for the global pair-wise sequence alignment in one dataset (*TFSPTRENR25*) (*Figure 7*).

V.B DISORDER PREDICTIONS

Prediction of intrinsic disorder in TF was performed using PONDR VL-XT [Li, X., *et al*, 1999; Romero, P., Z., *et al*, 1997 and 2001; http://www.pondr.com], CDF [Dunker AK, *et al* 2000], and Charge-Hydropathy Plots [Uversky, V. *et al*. 2000].

V.B.1 PONDR VL-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. The algorithm behind the predictors is a feed-forward neural network that uses sequence information from windows of generally 21 amino acids. Attributes, such as the fractional composition of particular amino acids or hydropathy, are calculated over this window, and these values are used as inputs for the predictor. One of the PONDR predictors, the VL-XT predictor integrates three feed forward neural networks: the VL1 predictor from Romero *et al*.

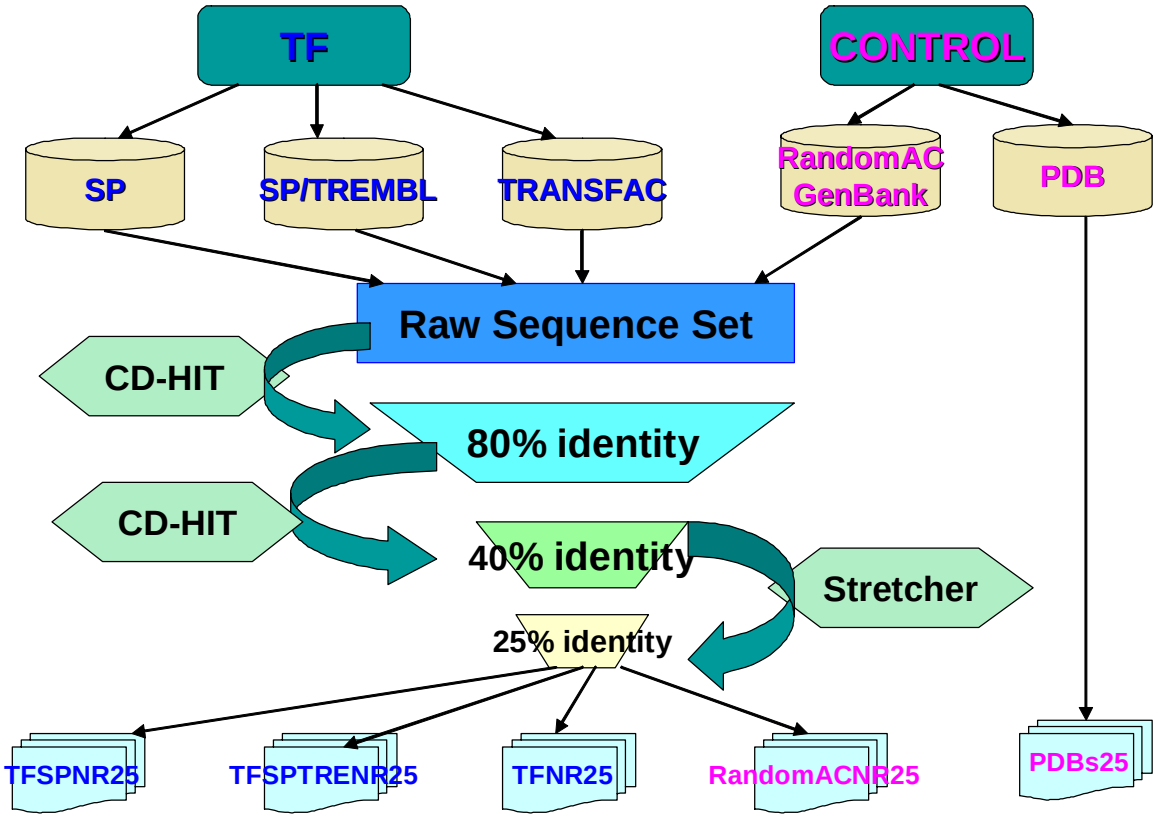


Figure 7: Flowchart for dataset construction and non-redundancy preparation

2000 and the N- and C- terminal predictors (XT) from Li *et al.* 1999. VL1 was trained using 8 disordered regions identified from missing electron density in X-ray crystallographic studies, and 7 disordered regions characterized by NMR. The XT predictors were also trained using X-ray crystallographic data. Output for the VL1 predictor starts and ends 11 amino acids from the termini. The XT predictors output 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 9 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 4 sequence positions.

V.B.2 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, so 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) [Sprent, P. 1993], can be calculated by determining the fraction of the distribution that lies below a given value [Dunker, *et al,* 2000; Oldfield CJ, *et al*, 2005]. On other hand, this method summarizes these per-residue predictions by plotting PONDR scores against their cumulative frequency, which allows ordered and disordered proteins to be distinguished based on the distribution of prediction scores.

V.B.3 Charge-Hydropathy Plots

Another established method of order-disorder classification is Charge-Hydropathy Plots [Uversky, V. *et al*. 2000]. Ordered and disordered proteins plotted in charge-hydropathy space can be separated to a significant degree by a linear boundary. The hydrophobicity of each amino acid sequence was calculated by the Kyte and Doolittle approximation within a window size of 5 amino acids. The hydrophobicity of individual residues was normalized to a scale of 0 to 1 in these calculations. The mean hydrophobicity is defined as the sum of the normalized hydrophobicities of all residues divided by the number of residues in the polypeptide. The mean net charge is defined as the net charge at pH 7.0, 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).

The computer programs of CDFs and Charge-Hydropathy Plots used in this project to classify proteins as completely disordered or completely ordered were written by Christopher J. Oldfield at Molecular Kinetics, Inc. For CDF, the program returns the value of all VL-XT based CDF bins for a protein and the classification of the protein based on a default 7 points scheme for both majority vote and unanimous methods. For Charge-Hydropathy Plots, the program will return the mean net charge, hydropathy, the formal euclidian distance, and class (ordered or disordered).

V.C TF DOMAIN INFORMATION

All the domain information was extracted from the section of 'Feature Table Data' in each entry in "*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 allocated to each of the data items within each FT line are shown in *table 1* (column numbers not referred to in the table are always occupied by blanks).

An example of a feature table for human CREB1 is shown in *table 2*.The residue positions (shown in *table 1* as 'From' endpoint and 'To' endpoint in Feature Table respectively) for each feature were used to retrieve the corresponding domain sequence

Table 1: the 'Feature Table'

Columns	Data item
1-2	FT
6-13	Key name
15-20	'From' endpoint
22-27	'To' endpoint
35-75	Description

Table 2: The feature table for human CREB

```
FT   Key Name     From   To        Description

FT   DOMAIN       101    160       KID.
FT   DNA_BIND     284    305       Basic motif.
FT   DOMAIN       311    332       Leucine-zipper.
FT   MOD_RES      133    133       Phosphoserine.
FT   MOD_RES      142    142       Phosphoserine (By similarity).
FT   VARSPLIC      88    101       Missing (in isoform CREB-B).
FT                                 /FTId=VSP_000596.
FT   CONFLICT       4      4       E -> D (in Ref. 5).
FT   CONFLICT       8      8       E -> D (in Ref. 5).
FT   CONFLICT     160    160       T -> A (in Ref. 5).
FT   CONFLICT     167    167       T -> A (in Ref. 5).
FT   CONFLICT     169    169       T -> A (in Ref. 5).
FT   CONFLICT     176    176       Q -> R (in Ref. 5).
FT   CONFLICT     184    184       A -> T (in Ref. 5).
FT   CONFLICT     188    188       G -> R (in Ref. 5).
FT   CONFLICT     195    195       N -> S (in Ref. 5).
FT   CONFLICT     210    210       T -> A (in Ref. 5).
```

(Swiss-Prot user manual http://au.expasy.org/sprot/userman.htm) and VL-XT PONDR prediction score for the local disordered prediction and the amino acid composition calculation.

V.D AMINO ACID COMPOSITION PLOTS

To compare the compositions of three different TF datasets with 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 TF are depleted compared with control in the indicated amino acids, and positive peaks indicate the reverse.

VI. RESULTS AND DISCUSSION

VI.A DATASET CHARACTERIZATION

The overview of five datasets used in this study is shown in *table 3 and 4, and figure 8*: *TFSPTRENR25* is a non-redundant, representative database from SWISS-PROT

Table 3: Dataset construction

DATASET	Source	No. Entries	NR80	NR40	NR25	Strip-off %
TFSPTRE	Swiss-Prot & TrEMBL	7195	4454	2633	1851	74%
TFSP	Swiss-Prot	2683	1827	1236	1082	60%
TF	TRANSFAC	1186	772	551	460	61%
RandomAC	NCBI (randomized AC)	2387	2314	2105	1935	19%
PDBs25	Dunker's Lab				1771	

and TrEMBL. This set contains 1851 sequences where no two sequences have more than 25% sequence identity after a serial sequence redundancy reduction from initially 7195 sequences retrieved. *TFSPNR25* contains 1082 representative TF sequences and is similar to *TFSPTRENR25* after serial sequence redundancy reductions as described in the Material and Methods section. The difference from *TFSPTRENR25* is that all entries in

the latter set were retrieved from *Swiss-Port* only, and feature table data (domain information) for every entry has been well annotated and defined in Swiss-Prot format. *TFNR25* was constructed based on the availability of both sequence and feature table data (domain information). 1186 TF sequences were initially retrieved from *TRANSFAC®*, which covers only eukaryotic *cis*-acting regulatory DNA elements and trans-acting factors. The final set contains 460 non-redundant sequences whose homology is less

Table 4: Description of Five Non-redundant TF Datasets (NR25)

Database Name	No. proteins	Min. protein length	Max. protein length	Average length	Median length
TFSPTRENR25	1851	31	3859	454	364
TFSPNR25	1082	53	6758	568	465
TFNR25	460	109	3759	549	442
RandomACNR25	1935	31	5038	480	364
PDBs25	1771	31	1235	173	135

NR25 / s25: no two proteins have sequence similarity higher than 25% identical residues for aligned subsequences

than 25%. *PDBs25* is well known as a set of non-homologous proteins and no two proteins have sequence similarity higher than a certain cutoff (25% identical residues for aligned subsequences), yet all structurally unique protein families are represented. As the main database of experimentally characterized structural information, Protein Data Bank (PDB) [H.M. Berman *et al* 2000] contains more than 20,000 structures of proteins,

nucleic acids and other related macromolecules characterized by methods such as X-ray diffraction and nuclear magnetic resonance (NMR) spectroscopy. However, current information in PDB is highly biased in the sense that it does not adequately cover the whole sequence/structure space. Evidence that trans-membrane, signal, disordered, and low complexity regions are significantly underrepresented in PDB has been reported reflected to the nature of the structural database of PDB [Peng K *et al* 2004]. To avoid this bias, we constructed a database called *randomACNR25*. We started with 4000 computer generated and randomized NCBI accession numbers and 2387 sequences were retrieved from Genbank. 1935 sequences survived after removing the sequence redundancy to less than 25%.

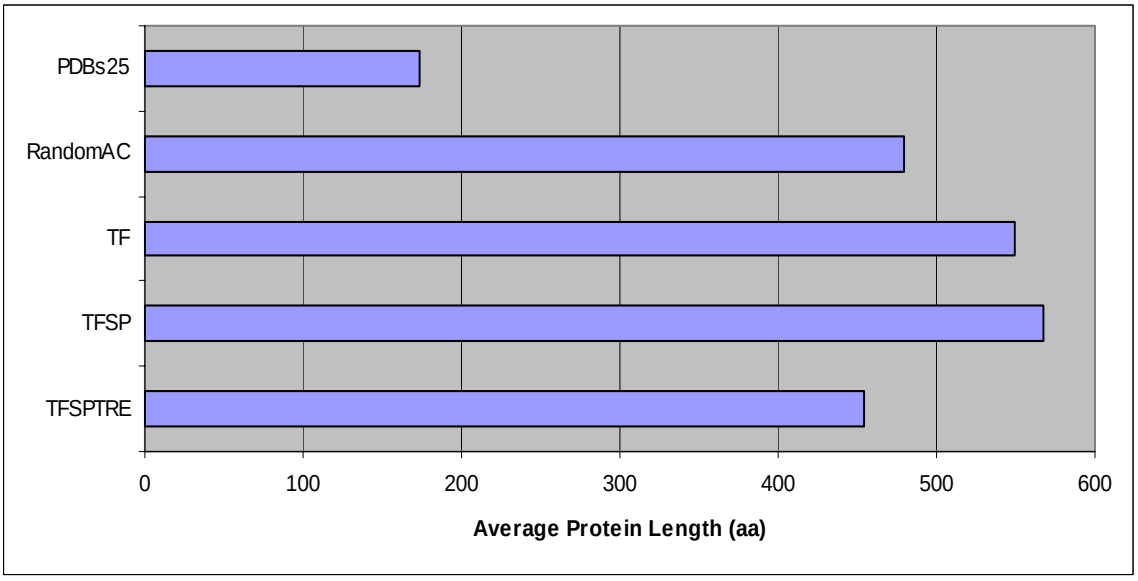


Figure 8: Average Protein Length (a. a.) in Five Datasets (NR25)

To better understand the degree of sequence redundancy in each preliminary dataset, one measure 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, we notified that the *randomACNR25* dataset has much lower *strip-off rate* (18.9%) compare to other three TF datasets. The rate for *TFSPTRENR25, TFSPNR25*, and *TFNR25* are 74.3%, 59.7%, and 61.2%, respectively (*Table 3 and 4*). Perhaps the constitution of sequence homology in the different datasets is the major contributor to the difference in the *strip-off* rate. As we described in the Material and Methods section, the sequences in *randomACNR25* set was retrieved by randomized accession number and resulted in a wide range of diversity. In contrast, all sequences in other three TF datasets are TFs, and some of them have probably been derived from same superfamily and have similar highly conserved motifs such as the zinc-finger, leucine-zipper, and homeobox, *etc*.

VI.B DISORDER PREDICTION ON TFS

To test for a generalized prevalence of intrinsic disorder in transcriptional regulation, we first used the Predictor Of Natural Disorder Regions (PONDR VL-XT) to

Table 5: Disorder prediction result on five datasets by VL-XT

	Num. of sequences	Average num. residues	Average num. Disorder res.	Overall% disorder	DO Score
PDBs25	1583	173.69	39.89	23%	0.31
RandomACNR25	1930	480.09	146.22	30%	0.33
TFNR25	460	549.19	283.73	52%	0.52
TFSPNR25	1080	568.20	263.91	46%	0.46
TFSPTRENR25	1819	454.21	216.55	48%	0.47

systematically analyze the intrinsic disorder in the three TF datasets. As shown in *Table 5*, the PONDR VL-XT predictions demonstrate that predicted disorder followed the ranking *TFNR25* > *TFSPTRENR25* > *TFSPNR25* > *RandomACNR25* > *PDBs25*. The difference between *TFSPNR25* and *TFSPTRENR25* is near negligible (46% and 48% for overall disorder; 0.46 and 0.47 for average disorder score, respectively) compared to the differences among other sets. The same ranking was observed in *Figure 9* when the percentages of TF with 30 or more consecutive disordered residues were calculated using two methods (*Figure 9* (**a**) and (**b**)). The percentages for consecutive disordered regions

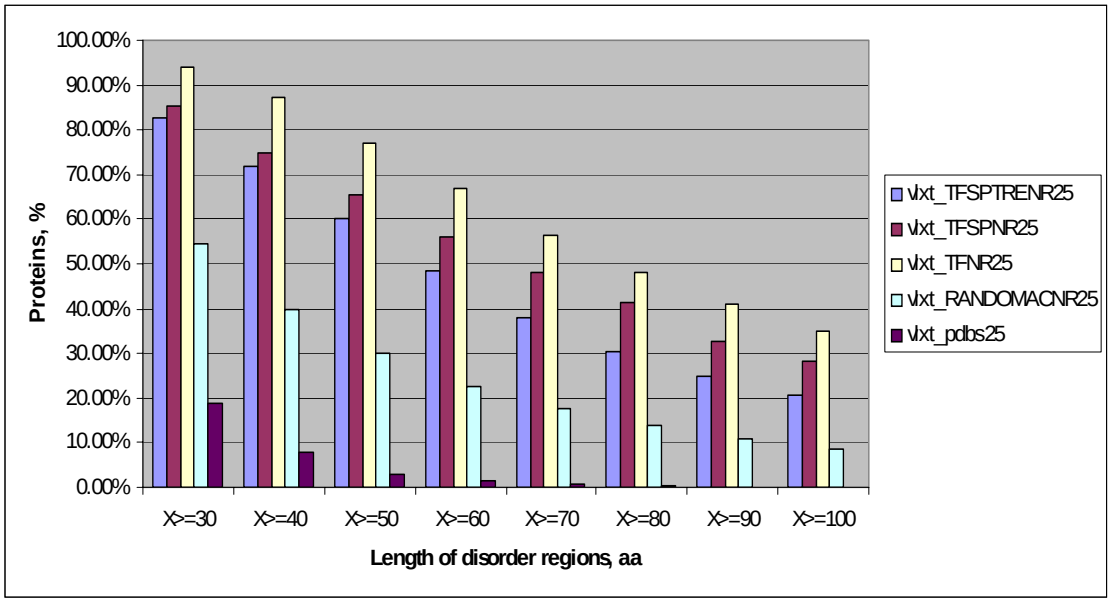


Figure 9 (a): Percentages of proteins in five datasets with>=30 to >= 100 consecutive residues predicted to be disordered

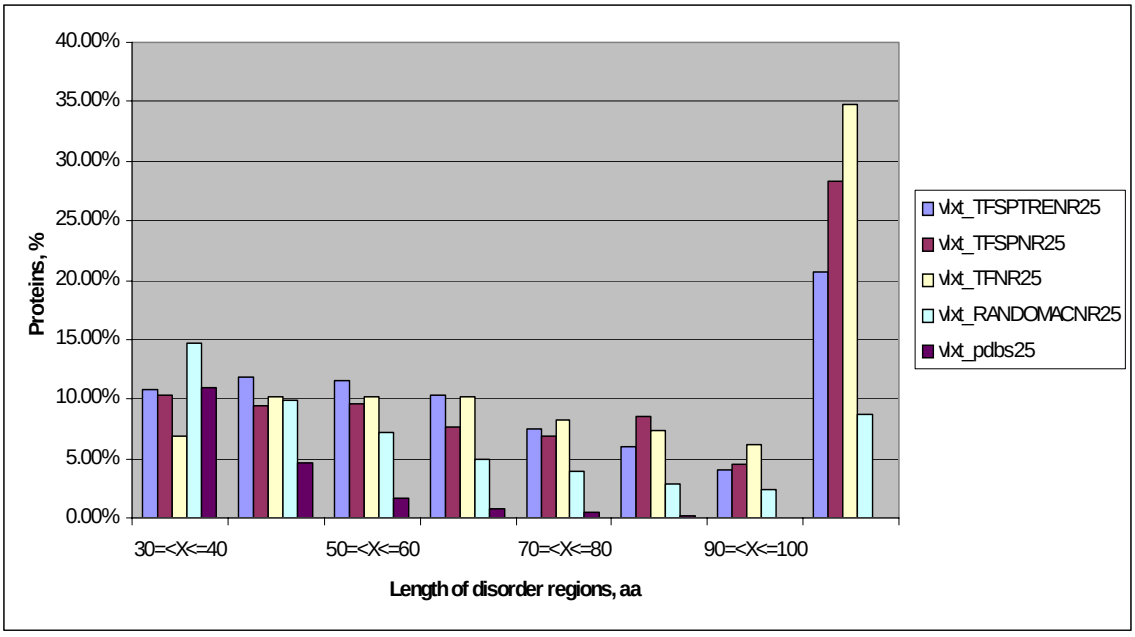


Figure 9 (b): Percentages of proteins in five datasets with consecutive residues Between 30 and 40; 40 and 50; … to >= 100 predicted to be disordered.

of >=30 residues in five datasets, *TFNR25, TFSPNR25, TFSPTRENR25*, *RandomACNR25*, and *PDBs25*, were 94.13%, 85.19%, 82.63%, 54.51%, and 18.64%, respectively (*Figure 9(a)*). All three TF datasets, *TFNR25, TFSPNR25,* and *TFSPTRENR25* have 5.0, 4.6, 4.4 fold more predicted disordered regions of >=30 consecutive residues as compare to *PDBs25,* respectively. When we compared to *RandomACNR25*, the increases for >=30 consecutive disordered residues in TF datasets, *TFNR25, TFSPNR25,* and *TFSPTRENR25,* were smaller. It is 1.7, 1.6, and 1.5, respectively. The nature of sequence collection and representation in these control sets may explain the fold differences for the disorder prediction. As described in the Materials and Methods, most proteins in PDB are ordered or partially ordered; in contrast, the

RandomACNR25 covers much wide range of sequence diversity, and have many more disordered proteins in the set and more closer to the nature of random sequence space.

As shown in *Figure 9 (a),* the percentages of the proteins that have disordered regions of >= 100 residues in five datasets, *TFNR25, TFSPNR25, TFSPTRENR25, RandomACNR25*, and *PDBs25*, is 34.78%, 28.24%, 20.62%, 8.65%, and 0.06%, respectively. TF datasets have significant higher percentage of proteins that have long disordered regions (>=100 residues) in comparison to *RandomACNR25* and *PDBs25*.

The comparison among the three TF datasets suggested that *TFNR25* set was enriched in predicted disorder than the other two TF datasets, *TFSPNR25* and *TFSPTRENR25,* although the difference between the later two sets is insignificant. We believe that *TFNR25* constructed from *TRANSFAC*® is a dataset that covers only eukaryotic TFs, and recent evidence suggests that eukaryotes have high percentage of native disorder than others [Dunker *et al* 2000].

Surprisingly, on or above 70% of the proteins in the three TF datasets are predicted to be wholly disordered by CDF (*Figure 10*), and similar degree of disorder was observed when the prediction was performed with PONDR VL-XT based on the single residue disorder score. It is a significant amount of wholly disordered proteins in the TF datasets compared to the two other control sets, 28.81% for *RandomACNR25* and 17.75% for *PDBs25*. Disorder enables complexes with low affinity coupled with high specificity and also facilitates the binding of one molecule to many partners [Dyson and Wright 2005]. These two characteristics may help us to understand why a large amount of wholly disordered proteins exist in the TFs data set as it might relate to the great need for control and regulation of gene transcription by these proteins.

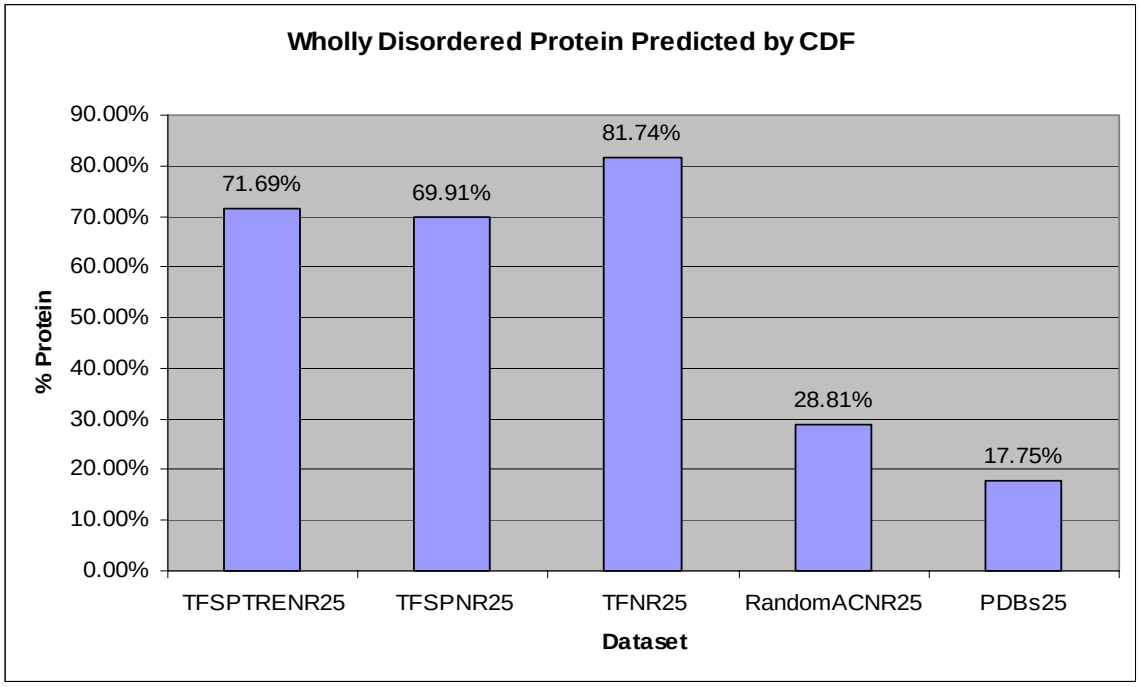


Figure 10: Prediction of Wholly Disordered TFs by Cumulative Distribution Functions (CDF)

We also noted that there are 17.75% proteins predicted to be wholly disordered in *PDBs25*. Wholly disordered proteins should not be expected to form crystals, so any such proteins predicted to be wholly disordered in PDB would be candidates for prediction errors. However, predicted disordered proteins in PDB might not all associated to prediction errors. There are several exceptions: (1) Sometimes fragments of proteins sometimes rather than whole proteins are crystallized; (2) Many intrinsically disordered proteins become ordered upon binding to partners. Such proteins can appear in PDB as ordered because the complex, not the individual protein, has been crystallized; (3) Proteins in PDB may contain segments of disorder that are associated with the putative wholly disordered proteins; (4) As shown in *table 4* and *figure 8*, the average size of proteins in PDBs25 set is 173 residues and most are much smaller fragments compared to

other datasets. It has been reported that small fragments could not be predicted well using PONDR VL-XT [Romero, P et al, 2001].

To apply the Charge-Hydropathy Plots for disorder prediction, the mean net charge and the mean normalized Kyte-Dollittle hydropathy were calculated for each protein in all five datasets, and the optimal boundary between the ordered and disordered proteins in charge-hydropathy space was determined by the procedure described. The percentage of wholly disordered proteins in three TF datasets is 28.70% for *TFNR25*, 23.80% for *TFSPNR25*, and 26.06% for *TFSPTRENR25* (*Figure 11*). The percentages are much smaller than that predicted by VL-XT, but it is still substantially higher in comparison with *RandomACNR25* (10.36%) and *PDBs25* (16.99%). Comparing the percentage of wholly disordered proteins in all five datasets predicted by CDF and

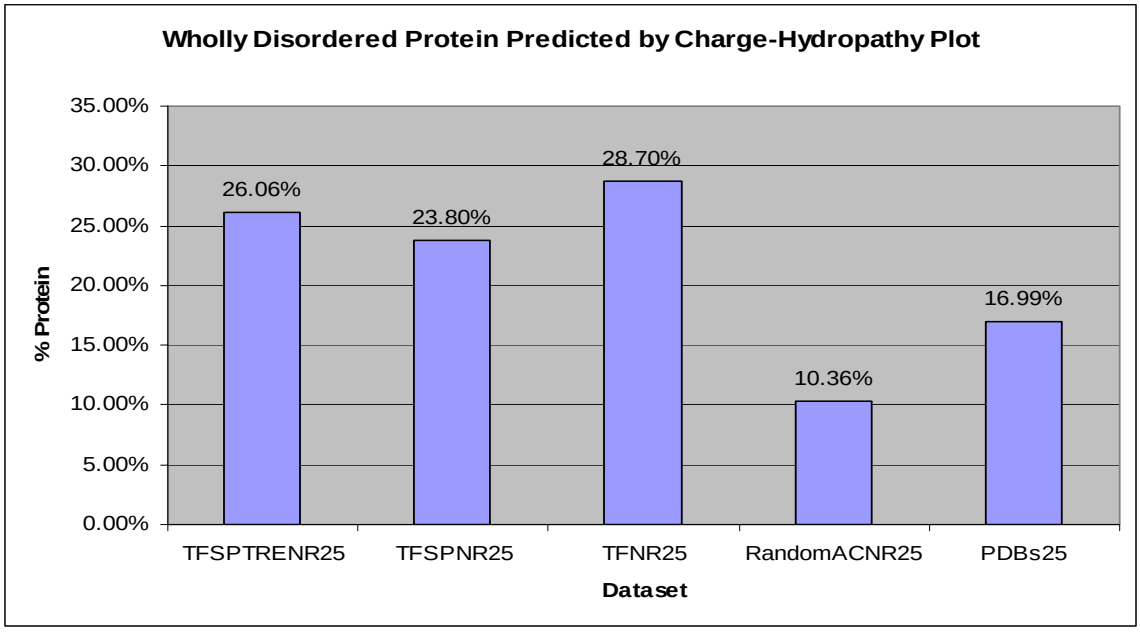


Figure 11: Prediction of Wholly Disordered TF by Charge-hydropathy Plots

Charge-Hydropathy methods, we found that the discrepancy in all datasets (except *PDBs25*) caused by the two different predictors was sizeable. The wholly disordered rate in all 4 datasets (*TFNR25, TFSPNR25, TFSPTRENR25,* and *RandomACNR25)* predicted by CDF was much higher (from 2.78 - to 2.94) than that predicted by Charge-Hydropathy plots. Surprisingly, the wholly disordered percentages in *PDBs25* predicted by these two methods are almost the same (17.75% vs 16.99%). This different magnitudes of disorder predicted between these two methods is similar to other published estimates that CDF analysis predicts from 1.2 – to 2.2 – fold more sequences to be disordered than charge-hydropathy [Oldfield, CJ *et al*, 2005]. As argued in these reports the difference in predictions by these two classifiers may be physically interpretable, in terms of the protein trinity model [Dunker, AK and Obradovic, Z.: 2001] or related protein quartet model [Uversky, V. N 2002]. On other hand, these two predictors probably caught two different stages (model) of proteins for their prediction. Proteins predicted to be disordered by the Charge-Hydropathy approach are likely to belong to the extended disorder class, on other hand, PONDR-based approaches can discriminate all disordered conformation (coil-like, premolten globules and molten globules) from rigid well-folded proteins, suggesting that charge-hydropathy classification is roughly a subset of PONDR VL-XT, in both predictions of disorder and feature space.

Regardless of the differing degree of disorder between these two prediction methods, both approaches predicted TF datasets to contain a significant higher portion of disordered proteins than either of the two control sets. It confirms the trends discovered by several individual case studies on a small number of TFs crystallized to date [Dyson HJ and Wright PE. 2005].

VI.C TF COMPOSITIONAL SPECIFICITY

The residue-wise composition of the TF sets (*TFSPTRENR25, TFSPNR25,* and *TFNR25*) and the two control datasets (*RandomACNR25* and *PDBs25*) were calculated to assess any specificity present therein. To visualize the differences between TF datasets and controls, the relative compositions were calculated as described by Romero *et al*, 2001. The amino acids in *Figure 12* (*a*) and (*b*) are arranged from the most rigid at the left to the most flexible at the right according to the scale of Vihinen *et al*. This scale is based on the averaged B-factor values for the backbone atoms of each residue type as estimated from 92 proteins. As the developers of this 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. The

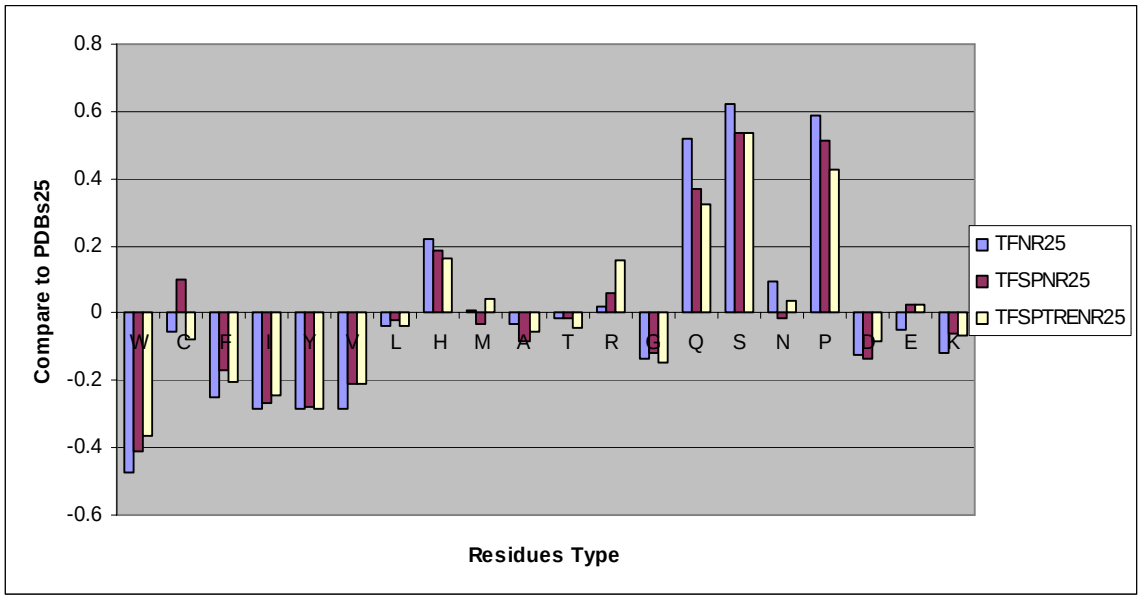


Figure 12 (a): Amino Acid Compositions of TF compared to PDBs25

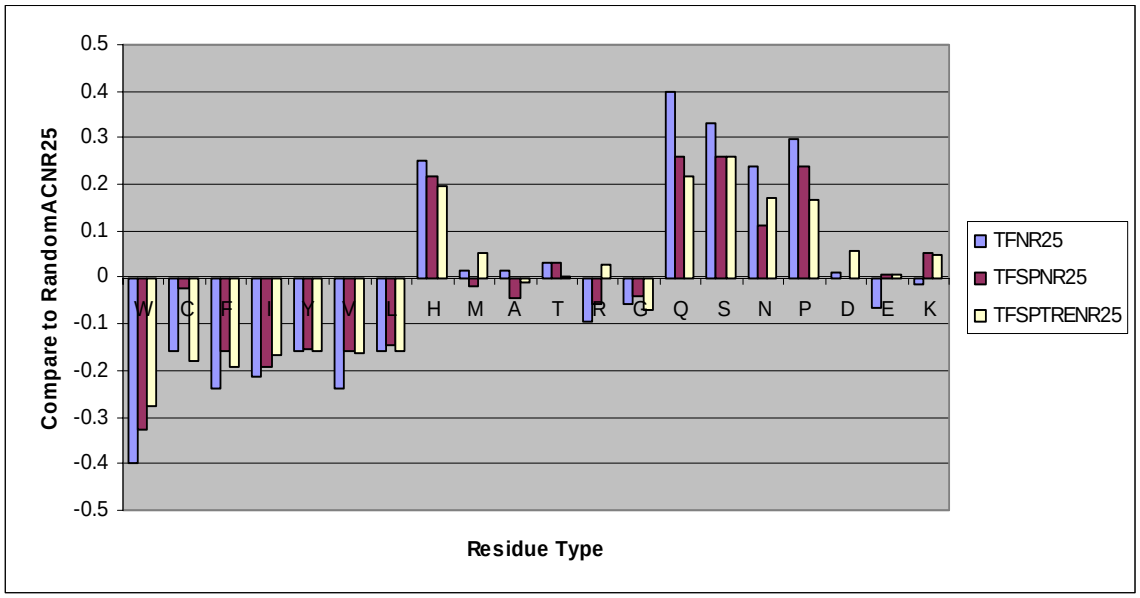


Figure 12 (b): Amino Acid Compositions of TF compare to RandomACNR25

amino acids to the left have been called *order-promoting* and those to the right *disorder-promoting* [Dunker *et al* 2001].

Overall, the amino acid compositions of *TFNR25, TFSPNR25,* and *TFSPTRENR25* are similar to each other but different from the two control sets. Like most other intrinsically disordered proteins [Dunker *et al* 2001, Romero *et al* 2001], all three TF datasets are substantially depleted in ***W, F, I, Y***, and ***V*** from 20% to 50% (shows as the negative peaks in *Figures 12(a) and 12 (b)*), and significantly enriched in ***Q, S***, and ***P*** from 30% to over 60% (*e.g*. the positive peaks for Q, S, and P in *Figure 12(a) and 12 (b)*) with only a few exceptions (especially ***H*** at left side, and ***G***, and ***N*** at right side). Obviously, the results of the composition for all three TF datasets not only exhibit most of the amino acid compositional bias presented by disordered proteins, but also reflect the

signature of TFs for DNA and protein binding. For example, ***H*** and ***C*** are highly over-represented in TF datasets because nearly a half of TFs contain several zinc-fingers [Tuple *et al* 2001] and one popular type of Zinc-finger is ***C2H2***. Zinc-finger has been characterized to play an important role in the overall recognition of DNA targets [Wolfe *et al* 1999]. Additionally, high occurrence of proline and glutamine in these proteins also suggests that these residues may contribute to conformational flexibility needed during the process of co-activators or repressors binding in transcriptional activation.

VI.D DISORDER IN TF DOMAIN AND SUBDOMAIN

To better understand and gain insight into the association between TF functions (DNA binding and transcriptional regulation) and their intrinsically disordered or ordered structure, we systematically dissected various annotated domains in one of the TF datasets (*TFSPNR25*) to analyze the region of disorder or order coupling with established function. The domain annotation was extracted from *TFSPNR25* dataset in the Swiss-Prot format. The disorder predictions were calculated, and grouped based on the 'key name' combined with 'description' in corresponding feature table as described in the Materials and Methods section. One may notice that some subdomain (motif) name is different from the list of 53 DNA-binding domains in Pfam models reported by several groups [Zupicich J *et al*, 2001; Stegmaier P. *et al* 2004]. This is because of the varied methods for domain annotations and classifications used in Swiss-Prot.

Table 6 shows the intrinsic disorder prediction with PONDR VL-XT on DBDs. From this table, we observed that *Basic Domains* was one of the most popular motifs among the DBDs of *TFSPNR25*. Its average length is 18.1, and the shortest and longest

Table 6: The Order/Disorder Region in TF DBD

Motif Name	Num. of Motif	Average Length	Shortest	Longest	Disorder Residues	Overall Disoder	Disorder Score
DNA-BIND	8	81.38	1	111	17.75	21.81%	0.3061
DNA_BIND: A.T hook	19	12.21	9	15	12.11	99.14%	0.9703
DNA_BIND: AP2/ERF	12	58.42	58	59	25.75	44.08%	0.4947
DNA_BIND: Basic motif	98	18.12	4	33	17.51	96.62%	0.9156
DNA_BIND: By-similarity	20	100.20	7	191	16.50	16.47%	0.2416
DNA_BIND: Copper-fist	4	40.00	40	40	12.75	31.88%	0.3624
DNA_BIND: CUT	7	87.86	87	88	38.14	43.41%	0.4517
DNA_BIND: DM	3	47.33	47	48	19.33	40.85%	0.4450
DNA_BIND: DNA-binding motif	2	23.50	10	37	0.00	0.00%	0.1035
DNA_BIND: ETS	14	82.00	81	85	17.43	21.25%	0.2748
DNA_BIND: Fork-head	20	92.75	87	97	16.00	17.25%	0.2238
DNA_BIND: H-T-H motif (potential	19	20.05	19	22	7.58	37.80%	0.4177
DNA_BIND: HMG box	39	70.33	65	114	32.41	46.08%	0.4504
DNA_BIND: Homeobox	91	60.73	56	81	28.42	46.80%	0.4897
DNA_BIND: Mef2-type (potential)	5	29.40	29	30	3.80	12.93%	0.1903
DNA_BIND: Myb	15	50.73	26	77	24.60	48.49%	0.4911
DNA_BIND: Nuclear-receptor-type	8	73.38	66	77	28.88	39.35%	0.3763
DNA_BIND: Potential	10	76.30	2	141	30.10	39.45%	0.4559
DNA_BIND: T-box	13	164.69	64	188	20.08	12.19%	0.1638
DNA_BIND: TF-B3	5	103.00	103	103	16.00	15.53%	0.2451
DNA_BIND: Tryptophan pentad re	5	102.20	101	103	28.60	27.98%	0.3163
DNA_BIND: WRKY	66	66.61	65	71	20.80	31.23%	0.3632
DNA_BIND: Zn(2)-Cys(6), fungal-t	13	28.62	27	31	13.77	48.12%	0.4572
ZN_FING: C2H2-type	1016	23.69	12	38	2.25	9.50%	0.1584
ZN_FING: PHD-type	29	52.97	44	62	4.34	8.20%	0.1337
ZN_FING: C4-type	21	24.10	18	36	6.33	26.28%	0.2593
ZN_FING: GATA-type	21	25.10	25	26	0.81	3.23%	0.0969
ZN_FING: C2HC-type	19	24.42	18	27	1.89	7.76%	0.1538
ZN_FING: RING-type	17	44.24	36	57	5.35	12.10%	0.1540
ZN_FING: B box-type	13	45.62	41	54	4.38	9.61%	0.1592
ZN_FING: MYM-type	9	40.67	35	59	0.22	0.55%	0.0799
ZN_FING: Zn-ribbon	9	26.44	23	39	0.11	0.42%	0.2034
ZN_FING: Dof-type	7	55.00	55	55	16.00	29.09%	0.3199
ZN_FING: MYND-type	3	37.00	37	37	1.67	4.50%	0.1099
ZN_FING: CHC2-type	2	23.00	23	23	0.00	0.00%	0.0111
ZN_FING: RanBP2-type	2	30.00	30	30	3.00	10.00%	0.1569
ZN_FING: U1-type	2	25.00	25	25	6.50	26.00%	0.2928
Average							0.3107

lengths are 4 and 33 residues, respectively. The average number of disordered residues is 17.5 and the overall percentage of disorder is significantly high, 96.6%. This data demonstrated that the region of basic motif in TF is highly unstructured or disordered. Our computational prediction has been strongly supported by following experimental data. In the early 90's, it was demonstrated that the basic regions of the bzip protein Fos and Jun [Patel, L *et al* 1990], C/EBP [Shuman J.D. *et al* 1990] and GCN4 [O'Neil, KT *et al* 1990] are unfolded off their DNA sites. The basic motifs were characterized by a large

excess of positive charges, preventing them from being structured when free in solution, but becoming α-helically folded when interacting with DNA [Weiss, M.A. *et al* 1990]. Usually, basic motif appears in tight connection with a dimerization domain like a leucine zipper (ZIP) helix (HLH). The flexible highly disordered basic region mediates the high-affinity and sequence-specific DNA binding after a prerequisite of dimerization via an induced-fit recognition of DNA [Anthony-Cahill SJ *et al* 1992; Drew H and Travers A. 1985; Ferre-D'Amare' AR *et al* 1994].

The results shown in *Table 6* demonstrated that *C2H2* zinc finger domain was not only the most prevalent, but also one of the highest ordered DNA-binding protein motifs among DBDs. Its overall percentage of disorder is as low as 9.5%. Five classes currently populate the super-class of zinc coordinating domains. The five classes are *C2H2, C4, DM, GCM* and *WRKY*. zinc finger motifs were originally identified as DNA binding structures in the RNA polymerase III transcription factor TFIIIA, which binds to the internal control region of the 5S RNA gene [J. Miller, *et al* 1985]. Since Wright and coworkers determined the first structure of an isolated zinc-finger domain by solution NMR [Lee MS, *et al* 1989], at least two types of zinc fingers (the classic 'zinc finger' proteins and the steroid receptors) have also been found in TFs that mediate transcription mediated by PolII. *C2H2* type of zinc finger motif is prevalent in the mammalian TFs and other higher or lower eukaryotes [Tuple, *et al* 2001] (*Figure 3*). It consists of average 24 amino acids (the shortest is 12, and the longest is 38 residues as shown in *Table 6*) with 2 cysteine and two histidine residues that bind zinc ion and folds the relatively short polypeptide sequence 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 [Wolfe SA *et al* 1999].

One of extremely important DBDs is *Helix-Turn-Helix domain*. There are five classes in the helix-turn-helix superfamily. It includes *Homeobox*, *Paired*, *Forked/winged helix*, *Tryptophan clusters*, and *Myb* domain. The analysis of PONDR VL-XT predictions on this class of protein domains demonstrates that predicted disorder followed the ranking *Myb>Homeobox>Tryptophan clusters> Forked/winged helix* (*Table 6*). The percentage of overall disordered residues is 48.49% 46.8%, 27.98%, 19.22%, and 17.25%, respectively.

The motif that has near 100% of overall disorder among the DBDs (*Table 6*) is the 'A-T hook'. AT-hook is a small DNA-binding protein motif that was first described in the high mobility group non-histone chromosomal protein HMGI/Y, and it preferentially binds to the narrow minor groove of stretches of AT-rich sequence. HMG-type and T-box are classified to be the group of *B-scaffold domains with minor groove contacts* according to this specific DNA binding feature. HMGI/Y was first identified by high electrophoretic mobility among the nuclear proteins, and participates in a wide variety of cellular processes including regulation of inducible gene transcription. Recent advances have contributed greatly to our understanding of how the HMGI/Y proteins participate in the molecular mechanisms underlying these biological events. 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.

Like many other eukaryotic proteins, most of zinc finger motifs act as independently folded globular domains (ββα) 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. To better understand the requirements for the linker and its role in DNA recognition, we dissected the linker region connecting two *C2H2* zinc fingers in *TFSPNR25* set to predict its intrinsic disorder, and to analyze the amino acid composition in this region. The results presented in *Table 7* indicate that

Table 7: Intrinsic Disorder in C2H2 linker

Length (res.)	number of linkers	total residues	disorder residues	Overall disorder	disorder score
linker =TGEKP	142	710	49	6.90%	0.0335
linker =5	518	2590	207	7.99%	0.0332
1<= linker <=10	659	3348	282	8.42%	0.1580
11<= linker <=50	96	2658	1179	44.36%	0.4226
51<= linker <=100	47	3464	2153	62.15%	0.5899
linker >100	76	18600	12180	65.48%	0.6300
total	878	28070	16394		

linker sequences between two *C2H2s* vary greatly in length and composition, but more than half of linkers (518 in length of five-residues out of total 878 in various lengths) 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 the fingers (141 out of 518 in

lengths of five residues) with this linker length have a consensus sequence of the form *TGEKP*. The PONDR prediction (6.9% for overall disorder) for *TGEKP* suggests that the *TGEKP* linker is actually well ordered. It has been reported that the TGEKP linker between fingers is flexible in the free protein in NMR studies, but becomes more rigid upon binding DNA (Foster MP, *et al*, 1997; Bowers PM *et al* 1999; Wuttke DS *et al* 1997). We also found that the magnitude of disorder increased as the length of linkers extended. The overall percentage of disorder is high as 65% when the linker length reaches 100 residues. The amino acid composition results also show that the zinc linkers not only have high occurrence of the five residues of the consensus sequence, *TGEKP*, but also show increased prevalence of some polar, uncharged amino acids (such as Ser and Gln), and are substantially depleted in *order-promoting* amino acids such as *W, C, F, I, Y, V, L*, and *N* (*Figure 13*). Conclusive evidence that the linker length and composition

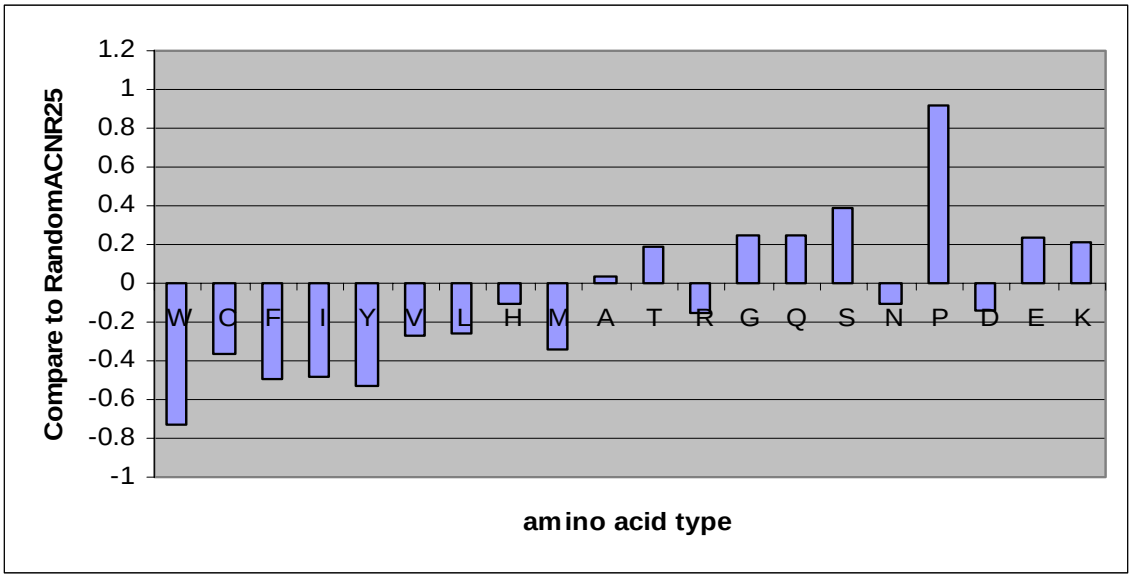


Figure 13: Amino Acid Composition for C2H2 Zinc-finger Linker

can influence both binding specificity and affinity, independently of the DNA-binding subdomains has come from several recently published studies [Leeuwen HC *et al* 1997; Peisach E and Pabo CO, 2003; Jantz D and Berg JM, 2004].

TF transcriptional activation depends on regions of as few as 30 to 100 amino acids that are separate from the DNA binding domain. So far, three different primary sequence motifs identified as the activation domains are *acidic*, *glutamine-rich*, and *proline-rich*. Deletion analyses of numerous transcription factors from mammals and Drosophila have identified several other classes that are rich in *serine* and *threonine* or other hydroxyl groups. However, some strong activation domains that are not particularly rich in any specific amino acid also have been identified. A few repression domains have also been identified; the best characterized is *alanine-rich*. Poor conservation of activation domain not only suggests that there are many targets and/or the interactions are generally non-specific, but also bring out the difficulty in completely covering the activation domain analysis in this study. Here we took only five different primary sequence motifs mentioned above for our study. The PONDR prediction indicates that the activation domain has highly tendency to be disordered (*Table 8*). The overall disorder is from 77% to 94%. The structure of not even one activation domain has yet been solved so far, although the 3-D structures of the DNA-binding domains from numerous eukaryotic transcription factors have been determined. This fact has indirectly confirmed our finding that most transcriptional activation domains are either unstructured or partly structured (*Table 7*). The intrinsically unstructured nature of these activation domains provides strong supporting evidence of a physiological role for coupled folding

Table 8: Disorder in TF Activation Domain

Domain Name	Num. of Domain	Shortest	Longest	Average Length	Disorder residue	Overall Disorder	Disorder score
acidic	81	6	213	40	31	77.48%	0.7485
alanine_rich	11	9	41	20	19	94.52%	0.8765
glutamine_rich	55	8	223	46	34	73.33%	0.7825
glycine_rich	16	8	76	29	26	88.98%	0.8205
proline_rich	30	11	248	61	48	77.65%	0.7609
serine_threonine_ rich	28	9	138	48	42	86.75%	0.8347
serine_rich	13	11	65	36	33	93.13%	0.8828
Average				**40**	**33**	**84.55%**	**0.8152**
(residue)							

and binding processes in transcriptional activation. Their inherent flexibility allows their local and global structure to be modified in response to different molecular targets, allowing one protein to interact with multiple cellular partners and allowing fine control over binding affinity.

VI.E TOP 15 PREDICTIONS OF DISORDERED TFS

To provide illustrative examples of novel predictions of disordered TFs, the 15 highest ranked TFs were then selected from a single organism, Homo sapiens, to avoid redundant orthologs in one of the TF datasets, *TFSPNR25* (*Table 8*). As a consequence of ranking TFs by overall percentage (above 80%) of disordered residues predicted by PONDR, these proteins represent extremes of long disordered regions and high disordered scores. It is surprising to find that four wholly unfolded native proteins (*HMGI-14, HMCI-C, SOX-15*, and *SOX-3*) among these top 15 disordered TFs belong to

one superfamily, *High Mobility Group (HMG)*. HMG is composed of three different families that have recently been renamed HMGA (a.k.a. HMGI/Y), HMGB (a.k.a. HMG-1 and -2) and HMBN (a.k.a. HMV-14 and -17) [Bustin, 2001]. HMG proteins are the founding members of a new class of regulatory elements called 'architectural

Table 9: Top 15 Human TFs from TFSPNR25 with >80% of Residues Predicted to be Disordered

			VL_XT				CDF			Charge_hydropathy			
AC.	Gene Code	TF name	length	Disordered Res.	Overall disordered	Longest D.	Count	All	Vote	Netcharge	Hydro-pathy	Class	exp. Confirmed
P05114	HMGN1; HMG14;	Nonhistone chromosomal protein HMG-14	99	99	100.00%	99	7	D	D	0.070707	0.32379	D	yes
P52926	HMGA2; HMGIC;	High mobility group protein HMGI-C	109	109	100.00%	109	7	D	D	0.110092	0.28787	D	yes
P23511	NFYA	CCAAT-binding transcription factor subunit B	347	334	96.25%	262	7	D	D	0.005764	0.4586	O	
Q02446	SP4	Transcription factor Sp4 (SPR-1)	784	706	90.05%	440	7	D	D	0.001276	0.45126	O	
P19532	TFE3	Transcription factor E3	743	656	88.29%	169	7	D	D	0.001346	0.43825	O	
P16220	CREB1	cAMP response element binding protein (CREB)	341	299	87.68%	212	7	D	D	0.014663	0.4478	O	yes
O60248	SOX15	SOX-15 protein	233	201	86.27%	105	7	D	D	0.051502	0.40629	D	yes
Q92766	RREB1	RAS-responsive element binding protein 1	755	648	85.83%	239	7	D	D	0.02649	0.39578	D	
P41225	SOX3	Transcription factor SOX-3	446	374	83.86%	115	7	D	D	0.03139	0.46776	O	yes
Q9ULX9	MAFF	Transcription factor MafF	164	137	83.54%	69	7	D	D	0.060976	0.4397	O	
P53567	CEBPG	CCAAT/enhancer binding protein gamma	150	125	83.33%	106	7	D	D	0.046667	0.39422	D	
P11831	SRF	Serum response factor (SRF)	508	423	83.27%	165	7	D	D	0.001969	0.4722	O	
Q02086	SP2; KIAA0048	Transcription factor Sp2	606	497	82.01%	243	7	D	D	0.051155	0.45928	O	
P16989	CSDA; DBPA	DNA-binding protein A	372	303	81.45%	115	7	D	D	0.032258	0.37724	D	
Q16520	BATF	ATF-like basic leucine zipper transcriptional factor B-ATF	125	101	80.80%	68	7	D	D	0.016	0.38213	D	

Res.: residues; **O**: ordered; **D**: Disordered

transcription factors' that participate in a wide variety of cellular process including regulation of inducible gene transcription, integration of retroviruses into chromosomes

and the induction of neoplastic transformation and promotion of metastatic progression of cancer cells [Reeves R. and Beckerbauer 2001]. 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 studies of a co-complex of individual AT hooks with a synthetic DNA substrate [J.R. Huth *et al* 1997], have elucidated the physical basis for 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 AT-hook DNA-binding characteristics, another principal reason why the HMGA proteins are able to physically interact with a large number of other proteins, most of which are TFs is because of the intrinsic flexibility associated with the unstructured properties. It has been reported that at least 18 different TFs so far were shown to be specifically associated with HMGA proteins as determined by various experimental methods (*Figure 14*). 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 extra- cellular signaling events and as integrators and effectors of the plethora of cellular responses to these stimuli [Reeves, 2001].

It is known that proteins, nucleic acids, and small molecules form a dense network of molecular interaction in a cell. Molecules are nodes of this network, and the interactions between them are edges. The architecture of molecular network can reveal important principles of cellular organization and function, similar to the way that protein

structure/unstructured tells us about the function and organization of a protein [Spirin V and Mirny LA 2003]. Regulation of gene expression involves a complex molecular

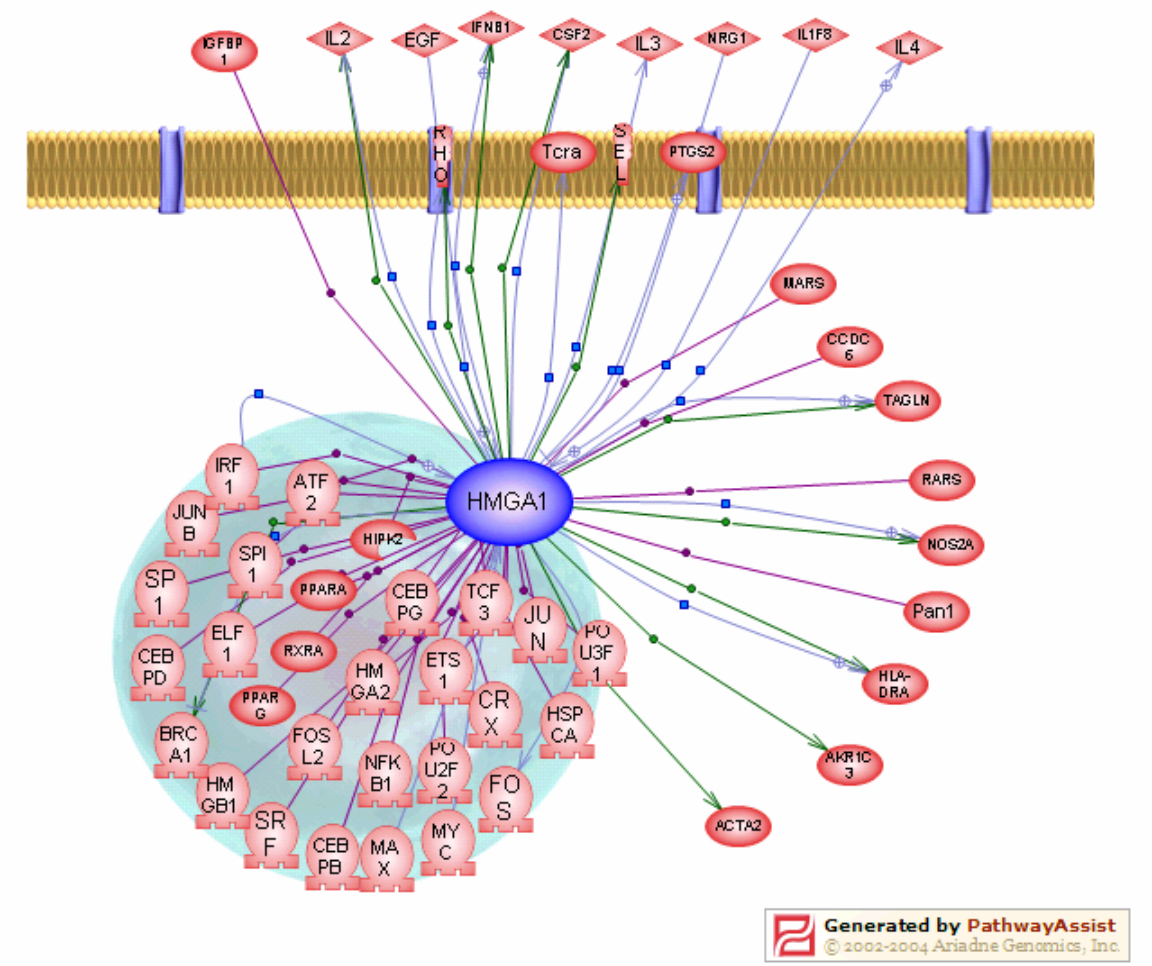


Figure 14: The HMGA1 protein acts as a 'hub' of nuclear function and interacts with at least 18 TFs in the nucleus.

network. DNA-binding transcription factors (TFs) are one of the important components in this network. Recently, it has been reported [Jeong *et al* 2000 and 2001] 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. Since one of the major functional advantages for intrinsically disordered proteins is the ability to bind to multiple different targets without sacrificing specificity to form the flexible nets [Dunker *et al*., 2005], and is responsible for the binding diversity of the broad cascade of protein-protein interactions, it is reasonable assume that disordered TFs are prime candidates for being essential protein 'hubs' for controlling many aspects of biological activity. The HMGA could be a typical example for this model.

Sox genes are a subgroup of specific HMG-box factors defined by similarity to Sry [Gubbay J, 1990], the mammalian testis-determining factor encoded by the Y chromosome, and are part of a larger family of transcription factors with DNA binding domains related to the general chromatin protein HMG1. Like other members in the HMG superfamily, the HMG-domain of Sox genes has interesting properties; it binds in the minor groove and induces a large bend in the DNA helix, prompting the suggestion that these proteins may have a chromatin architectural function. Unlike many members of the family, SOXs have restricted tissue specificity and exhibit a moderate degree of sequence specificity. Genetic analyses of *Sox* genes in humans, mice, and *Drosophila melanogaster* have demonstrated essential roles in specific cell fate decisions [Wagner T, *et al* 1994; Schilham MW, 1996]

So far, the free-solution structure of neither hSRY nor mLEF-1 has been determined. Using calorimetric measurements, the mSox-5 HMG box have shown that significant levels of protein refolding occur on association, in addition to the DNA bending [Crane-Robinson C *et al* 1998; Privalov *et al*. 1999].

The undergoing DNA-dependent order-disorder transition of Sox domains appears to play an important role for the adaptability of the motif's angular surface to enable a Sox protein to induce different architectures in different functional contexts. We can imagine that target genes for a given factor will differ, for example, in the precise sequence of Sox binding sites and its combinatorial relation to other factor-binding sites in the same promoter or enhancer. Because context-dependent changes in overall architecture may differentially affect transcription, a single factor may exert fine control over relative levels of expression within a set of target genes.

VI.F TF DISORDER IN DIFFERENT SPECIES

Table 10 lists top 11 popular species and its prediction of disorder in the *TFSPNR25* dataset. The results demonstrate that there are clear patterns show that TFs from eukaryotes have more intrinsic disorder than those from bacteria. TFs in eukaryotes are distinguished from these in bacteria by having the highest percentage of sequences predicted to have disordered segments >=50 in length: from 56% for Mouse-ear cress to 77% for human. In contrast, the percentage of sequences predicted to have disordered segments >=50 in length in prokaryotes ranged from 12.5% to 15%. One argument is whether the length of sequences contributes to the percentage difference between eukaryotes and bacteria since the average length of sequence in eukaryotes is longer than that in bacteria. The overall percentage of disordered, one measurement regardless of sequence length, indicates clearly that eukaryotic TFs have higher overall disordered rate than prokaryotic TFs. There is so far no conclusive evidence but there are some hints to explain why there is a large increase in intrinsic disorder for the eukaryotes. One of the explanations is that eukaryotes have well-developed elaborated gene transcription system,

and this system is in great need of TF flexibility. The intrinsically disordered TFs or partially unstructured regions can offer such important advantage in response to different

Table 10: TF Disorder in Different Species

Species Code	Common Name	Average Length	Shortest	Longest	Disorder Residues	Overall Disorder	L>=30	L>=40	L>=50	Num. of Proteins
HUMAN	Human	652.85	99	5262	325.81	49.90%	91.50%	83.28%	77.71%	341
MOUSE	Mouse	673.07	73	4903	337.56	50.15%	86.03%	82.35%	74.26%	136
YEAST	Yeast	565.19	103	1703	222.87	39.43%	85.07%	73.88%	59.70%	134
RAT	RAT	430.56	130	2148	206.42	47.94%	90.98%	73.77%	59.02%	122
ARATH	Mouse-ear cress	418.42	133	1895	195.15	46.64%	90.65%	71.96%	56.07%	107
DROME	Fruit fly	666.13	130	2065	361.72	54.30%	95.65%	91.30%	85.51%	69
BACSU	(Bacillus subtilis)	284.85	55	805	65.73	23.07%	40.00%	25.00%	15.00%	40
SCHPO	Fission yeast	474.73	163	979	194.61	40.99%	78.79%	66.67%	60.61%	33
CAEEL	C. elegans	488.04	142	1203	200.30	41.04%	82.61%	60.87%	56.52%	23
BRARE	Zebrafish	370.61	167	610	194.94	52.60%	83.33%	72.22%	61.11%	18
ECOLI	(Escherichia coli)	343.25	133	901	105.69	30.79%	62.50%	50.00%	12.50%	16

L: longest disorder region

molecular targets, allowing one protein to interact with multiple cellular partners and allowing fine control over binding affinity. In contrast, prokaryotes are subject to strong selective pressure on biochemical efficiency and do not have highly-regulated gene regulation system.

VII. CONCLUSIONS

Our results have demonstrated that the percentages of intrinsically disordered proteins in the three TF datasets were significantly higher than in the other two control sets. This prevalent phenomenon implies that intrinsic disorder in TFs may play a critical role in molecular recognition, DNA binding, and transcriptional regulation. Intrinsic disorder enables DNA-protein or protein-protein interaction with low affinity coupled with high specificity and also facilitates the binding of one molecule to many partners.

The amino acid compositions of the three TF datasets differ significantly from the two control sets. All three TF datasets are substantially depleted in *order-promoting* residues such as ***W, F, I, Y***, and ***V***, and significantly enriched in *disorder-promoting* ones such as ***Q, S***, and ***P***. The TF compositional specificity not only exhibits most of the amino acid compositional bias presented by disordered proteins, but also reflects the signature of TFs for DNA and protein binding.

Disorder predictions on TF domains showed that the AT-hook and basic region of DBDs were highly disordered. The C2H2 zinc-fingers were predicted to be highly ordered; however, the longer the zinc finger linkers, the higher the predicted magnitude of disorder. Overall, the degree of disorder in TF activation regions is much higher than that in DBDs.

Our data confirmed that the degree of disorder was significantly higher in eukaryotic TFs than in prokaryotic TFs, and suggested that eukaryotes have well-developed elaborate gene transcription system, and this system is in great need of TF flexibility. The intrinsically disordered TFs or partially unstructured regions can offer

such important advantage in response to different molecular targets, allowing one protein to interact with multiple cellular partners and allowing fine control over binding affinity.

VIII. REFERENCES:

Ayed A, Mulder FA, Yi GS, Lu Y, Kay LE, Arrowsmith C. H. (2001) Latent and active p53 are identical in conformation. *Nat Struct Biol* 8:756-760.

Barker A. and Muller-Hill B: (1998) Towards a mutant analysis of the tertiary structures of functional DNA-binding motifs, 432:1-3

Bewley, C. A., Gronenborn, A. M. and Clore, G. M. 1998 Minor groove-binding architectural proteins: structure, function, and DNA recognition. Annu. Rev. Biophys. Biomol. Struct. 27,105 -131

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.

Bowers PM, Schaufler LE, Klevit RE. (1999) A folding transition and novel zinc finger accessory domain in the transcription factor ADR1. *Nat. Struct. Biol*. 6:478-85

Cai W.; Hu L,; Foulkes J.G. (1996) *Current Opinion in Biotechnology*, 7, 608-615.

Clark KL, Halay ED, Lai E, Burley SK. (1993) Co-crystal structure of the HNF-3/fork head DNA-recognition motif resembles histone H5. *Nature* 364:412–20

Crane-Robinson C, Read CM, Cary PD, Driscoll PC, Dragan AI, Privalov PL (1998) The energetics of HMG box interactions with DNA. Thermodynamic description of the box from mouse Sox-5 *J Mol Biol* 281:705–717

Dhawale and Lande (1994) *Nucleic Acids Res*. 21:5537-5546

Dunker, A. K., Obradovic, Z., Romero, P., Garner, E. C., and Brown, C. J. (2000) Intrinsic protein disorder in complete genomes *Genome Informatics*, 11, 161-171.

Dunker, A. K., and Obradovic, Z. (2001) The protein trinity-linking function and disorder. *Nat. Biotechnol* 19, 805-806

Dunker, A. K., Brown, C. J., Lawson, J. D., Iakoucheva, L. M., and Obradovic, Z. (2002) Intrinsic Disorder and Protein Function, *Biochemistry* 41, 6573-6582.

Dunker, A. K., Cortese M. S., Iakoucheva, L. M., and Uversky V. 2005 Flexible Nets: The roles of intrinsic disorder in protein interaction networks, *submitted*

Dyson, H. J., and Wright, P. E. (2002) Coupling of folding and binding for unstructured proteins *Curr. Opin. Struct. Biol* 12, 54-60

Dyson, H. J., and Wright, P. E. 2005 Intrinsically Unstructured Proteins and Their Functions, *Nature Review: Molecular Biology* 6:197-208

Faisst and Meyer. 1992 *Nucleic Acids Res*. 20:3-26

Frankel AD and Kim PS: 1991 *Cell*, 65:717-719

Foster MP, Wuttke DS, Radhakrishnan I, Case DA, Gottesfeld JM, Wright PE 1997 Domain packing and dynamics in the DNA complex of the N-terminal zinc fingers of TFIIIA, *Nat. Struct. Biol*. 4:605-8

Grossmann JG, Sharff AJ, O'Hare P, Luisi B: 2001 Molecular shapes of transcription factors TFIIB and VP16 in solution: implications for recognition. *Biochemistry* 40:6267-6274.

Gubbay J, Collignon J, Koopman P, Capel B, Economou A, Munsterberg A, Vivian N, Goodfellow P, Lovell-Badge R 1990 A gene mapping to the sex-determining

region of the mouse Y chromosome is a member of a novel family of embryonically expressed genes. *Nature* 346:245–250

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

H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N.Shindyalov and P.E. Bourne 2000 The Protein Data Bank, *Nucleic Acids Res.*, 28, 235.

Hobohm U. et al. 1992 Selection of representative protein data sets. *Prot. Sci.* 1, 409-417

Jones DT, Ward JJ 2003 Prediction of disordered regions in proteins from position specific score matrices, *Proteins* 53 (suppl 6): 573-578.

Kissinger CR, Liu B, Martin-Blanco E, Kornberg TB, Pabo CO. 1990 Crystal structure of an engrailed homeodomain-DNA complex at 2.8 ° A resolution: a framework for understanding homeodomain-DNA interactions *Cell* 63:579–90

Klemm JD, Rould MA, Aurora R, Herr W, Pabo CO. 1994 Crystal structure of the Oct-1 POU domain bound to an octamer site: DNA recognition with tethered DNA-binding domains. *Cell* 77:21–32

Laity JH, Dyson HJ, Wright PE: 2000 DNA-induced a-helix capping in conserved linker sequences is a determinant of binding affinity in Cys2-His2 zinc fingers. *J Mol Biol* 295:719-727

Laity JH, Dyson HJ, Wright PE: 2000 Molecular basis for modulation of biological function by alternate splicing of the Wilms' tumor suppressor protein, *Proc Natl Acad Sci USA* 97:11932-11935

Latchman D.S. 2000 Transcription Factors as Potential Targets for Therapeutic Drugs. *Current Pharmaceutical Biotechnology*, July 2000, vol. 1, no. 1, pp. 57-61(5)

Latchman D.S. 1998 Eukaryotic transcription factors, Third Edition, Academic Press, London, San Diego

Latchman D.S. 1996 *New England Journal of Medicine* 334, 28-33

Li, W., Jaroszewski, L. and Godzik, A. 2001 Clustering of highly homologous sequences to reduce the size of large protein database, *Bioinformatics*, 17, 282-283

Li, W., Jaroszewski, L. and Godzik, A. 2002 Sequence clustering strategies improve remote homology recognitions while reducing search times; *Protein Engineering*, Vol. 15, No. 8, 643-649

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

Linding R, Russell RB, Neduva V, Gibson TJ 2003 GlobPlot: Exploring protein sequences for globularity and disorder, *Nucleic Acids Res* 31:3701-3708. [URL:http://globplot.embl.de/.]

Liu J, Tan H, Rost B 2002 Loopy proteins appear conserved in evolution. *J Mol Biol* 322:53-64.

Michael Levine, Robert Tjian 2003 Transcription regulation and animal diversity, *Nature* 424, 147 - 151 (10 Jul 2003) Review

Mitchell, P.J. and Tijan, R 1989 Transcriptional regulation in mammalian cells by sequence-specific DNA binding proteins, *Science* 245: 371-378

Ogata K, Hojo H, Aimoto S, Nakai T, Nakamura H 1992 Solution structure of a DNA-binding unit of Myb: a helix-turn-helix-related motif with conserved

tryptophans forming a hydrophobic core, *Proc. Natl. Acad. Sci. USA* 89:6428–32

Park J, Holm L, Heger A, Chothia C. 2000 RSDB: representative protein sequence databases have high information content. *Bioinformatics* 16(5): 458-64.

Patikoglou G and Burley, SK 1997 Eukaryotic Transcription Factor-DNA Complexes, *Annu. Rev. Biophys. Biomol. Struct*. 1997. 26:289-325

Peng K. Obradovic Z. Vucetic S. 2004 Exploring bias in the Protein Data Bank using contrast classifiers, *Pacific Symposium on Biocomputing*., : 435-46

Ptashne, M. 1988 How eukaryotic transcription factors work. *Nature* 335: 683-689.

Radhakrishnan I, Pérez-Alvarado GC, Parker D, Dyson HJ, Montminy MR, Wright PE: 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.

Reeves R. and Beckerbauer L 2001 HMGI/Y: flexible regulators of transcription and chromatin structure. *Biochimica et Biophysica Acta* 1519:13-29

Reeves R. and Beckerbauer L.; HMGI/Y proteins 1988 flexible regulators of transcription and chromatin structure, *Biochimica et Biophysica Acta* 1519 (2001) 13-29

Reeves R.: Molecular biology of HMGA proteins: 2001 hubs of nuclear function. *Gene* 277:63-81

Romero P, Obradovic Z, Kissinger CR, Villafranca JE, Dunker AK1997 Identifying disordered regions in proteins from amino acid sequences. *IEEE Int Conf Neural Netw* 1:90-95.

Romero P, Obradovic Z, Li X, Garner EC, Brown CJ, Dunker AK 2001 Sequence complexity of disordered protein. *Proteins* 42:38-48 [URL: http://www.PONDR.com]

Romero, P., Z. Obradovic, and A. K. Dunker, 1997 Sequence data analysis for long disordered regions prediction in the calcineurin family, *Genome Informatics*, 8:110-124

Romero, P., Z. Obradovic, X. Li, E. Garner, C. Brown, and A. K. Dunker 2001 Sequence complexity of disordered protein. *Proteins, Struct. Funct. Gen.*, 42:38-48.

Saudek V, Pastore A, Castiglion-Morelli MA, Frank R, Gausepohl H, et al 1991 The solution structure of a leucinezipper motif peptide, *Protein Eng.* 4:519–29

Schilham MW, Oosterwegel MA, Moerer P, Ya J, de Boer PA, van de Wetering M, Verbeek S, Lamers WH, Kruisbeek AM, Cumano A, Clevers H 1996 Defects in cardiac outflow tract formation and pro-B-lymphocyte expansion in mice lacking Sox-4, *Nature* 380:711–714

Spirin V. and Mirny LA: 2003 Protein complexes and functional modules in molecular networks, *PNAS* 100: 12123-12128

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

Steitz TA. 1990 Structural studies of protein-nucleic acid interaction: the sources of sequence-specific binding. Q. *Rev. Biophys.* 23:105–80

Tuple, R. Perini G, and Green MR.: 2001 Expressing the human genome, *Nature*, 409:832-833

Uversky, V. N. 2002 Natively unfolded proteins: a point where biology waits for physics *Protein Sci* 11, 739-756.

Uversky, V., Gillespie, J., and Fink, A. 2000 Why are natively unfolded proteins unstructured under physiological conditions? *Proteins* 41, 415-427

Uversky, V., Oldfield, C.J., and Dunker, A. K. 2005 Showing your ID: Intrinsic disorder as an ID for recognition, regulation, and cell signaling, *submitted*.

Villard J.2004 Transcription regulation and human diseases, *SWISS MED WKLY*, 134:571-579

Wagner T, Wirth J, Meyer J, Zabel B, Held M, Zimmer J, Pasantes J, Bricarelli FD, Keutel J, Hustert E, Wolf U, Tommerup N, Schempp W, Scherer G 1994 Autosomal sex reversal and campomelic dysplasia are caused by mutations in and around the SRY-related gene SOX9. *Cell* 79:1111–1120

Wingender, E. 1988 Compilation of transcription regulating proteins, *Nucleic Acids Res.* 16: 1879–1902

Wolfe SA. Nekludova L. and C.P. Pabo 1999 *Annu. Rev. Biophys. Biomol. Struct*. 3:183-212

Wuttke DS, Foster MP, Case DA, Gottesfeld JM, Wright P E. 1997Solution structure of the first three zinc fingers of TFI-IIA bound to the cognate DNA sequence: determinants of affinity and sequence specificity. *J. Mol. Biol*. 273:183-206

APPENDIX: CURRICULUM VITA

Jiangang (Al) Liu

5404 Alvamar Place, Carmel, IN 46033

Tel.: (317) 566-8879 (home); (317) 433-0033(lab.)

Fax: (317) 566-8879, E-mail: jliu1@iupui.edu

EDUCATION

M.S. (*will be graduated in June*) in Bioinformatics, IUPUI (2005)

B. S. (Equivalent) in Computer Science, University of Alabama at Birmingham (1998)

Ph.D. in Molecular Biology and Sport Medicine, Beijing Medical University, China (1990)

B. S. in Medicine, Shaoyang Medical College, China (1982)

EMPLOYMENT HISTORY AND EXPERIENCES

2002-present: Computational Biologist, Lilly Bioinformatics Group, Eli Lilly and Company

1999-2002: Biologist, Department of Cardiovascular Research, Eli Lilly and Company

1995-1999: Research Associate at University of Alabama at Birmingham.

1992-1995: Postdoctoral Fellow at University of Alabama at Birmingham.

1990-1991: Assistant Professor at Beijing Medical University, China.

PUBLICATIONS:

1. **Liu J**, Perumal N., Uversky V. and Dunker A. K. Intrinsic Disorder in Transcription Factors *In preparation.*
2. Liang JD, **Liu J**, McClelland P, Bergeron M.: Cellular localization of BM88 mRNA in paraffin –embedded rat brain sections by combined immunohistochemistry and non-radioactive hybridization. *Brain Res Res Protoc* 7(2): 121-30, 2001.

3. Mayne R. Ren ZX. **Liu J.** Cook T., Carson M. Narayana S. VIT-1: the second member of a new branch of the von Willebrand factor A domain superfamily, *Biochemical Society Transactions*, 27: 35192-9.
4. Paassilta, P., Pilhlajamaa, T., Annunen, S., Brewton, R.G., Wood, B.M., Johnson, C.C., **Liu, J.,** Gong, Y., Warman, M.L., Prockop, D.J., Mayne, R., and Ala-Kokko, L. Complete sequence of the 23-kilobase human COL9A3 gene: Detection of GLY-X-Y triplet deletions that represent neutral variants. *Journal of Biological Chemistry* 274, 22469-22475, 1999.
5. **Liu, J**., Swasdison, S., Xie, W., Brewton, R. G. and Mayne, R.: Primary structure and expression of a chicken laminin b chain: evidence for four b chains in birds. *Matrix Biology* 16: 471-481, 1998.
6. Li, F., **Liu, J**., Mayne, R., and Wu, C.: Identification and characterization of a mouse protein kinase that is highly homologous to human integrin-linked kinase. *Biochemica et Biophysica Acta* 1358: 215-220, 1997.
7. **Liu, J**., and Mayne, R.: The complete cDNA coding sequence and tissue-specific expression of the mouse laminin a4 chain. *Matrix Biology* 15: 433-437, 1996
8. **Liu, J**. Qu, M., Zhang, M., Li, W., and Guo, S.: Preparation and characterization of monoclonal antibody against rabbit collagen III. *J. Sports Med.* 13: 68 -74, 1994.
9. **Liu, J.,** and Qu, M.: Application of monoclonal antibodies to different collagen types in collagen immunolocalization. *J. Sports Med.* 11: 102-105, 1992.
10. **Liu, J**., Qu, M., Zhang, M., Ting, L., Liu, Y., Lin, Z. and Li, W.: The affect of monoclonal antibodies to collagen III on metabolism of human articular chondrocyte in vitro. *J. Sports Med.* 11: 195-197, 1992.
11. **Liu, J.,** Qu, M., Zhang, M., Lou, T., Liu, Y., Lin, Z., and Li, W.: The change of collagen I, II and III expression of human articular chondrocyte in vitro, *J. Sports Med.* 11: 6-8, 1992,
12. **Liu, J.** Qu, M., Zhang, M. and Liu, Y.: The alteration in expression of collagen I and III during repair of rabbit Achilles tendon, *J. Sports Med*. 11: 193-195, 1992.
13. **Liu, J.** and Gao, S.: Histological, histochemical study of myopathy induced by Dexamethasone, *J. China Med. Univ*. 19: 9-13, 1990