ARTICLE
Journal of Cellular Biochemistry 112:3256–3267 (2011)

Journal of Cellular Biochemistry

A Majority of the Cancer/Testis Antigens Are Intrinsically Disordered Proteins

Krithika Rajagopalan,[1] Steven M. Mooney,[1†] Nehal Parekh,[1†] Robert H. Getzenberg,[1,2,3] and Prakash Kulkarni[1,2*]

[1]*Department of Urology, James Buchanan Brady Urological Institute, Baltimore, MD 21287*
[2]*The Sidney Kimmel Comprehensive Cancer Center, Baltimore, MD 21287*
[3]*Pharmacology and Molecular Sciences, The Johns Hopkins University School of Medicine, Baltimore, MD 21287*

ABSTRACT

The cancer/testis antigens (CTAs) are a group of heterogeneous proteins that are typically expressed in the testis but aberrantly expressed in several types of cancer. Although overexpression of CTAs is frequently associated with advanced disease and poorer prognosis, the significance of this correlation is unclear since the functions of the CTAs in the disease process remain poorly understood. Here, employing a bioinformatics approach, we show that a majority of CTAs are intrinsically disordered proteins (IDPs). IDPs are proteins that, under physiological conditions in vitro, lack rigid 3D structures either along their entire length or in localized regions. Despite the lack of structure, most IDPs can transition from disorder to order upon binding to biological targets and often promote highly promiscuous interactions. IDPs play important roles in transcriptional regulation and signaling via regulatory protein networks and are often associated with dosage sensitivity. Consistent with these observations, we find that several CTAs can bind DNA, and their forced expression appears to increase cell growth implying a potential dosage-sensitive function. Furthermore, the CTAs appear to occupy "hub" positions in protein regulatory networks that typically adopt a "scale-free" power law distribution. Taken together, our data provide a novel perspective on the CTAs implicating them in processing and transducing information in altered physiological states in a dosage-sensitive manner. Identifying the CTAs that occupy hub positions in protein regulatory networks would allow a better understanding of their functions as well as the development of novel therapeutics to treat cancer. J. Cell. Biochem. 112: 3256–3267, 2011.

KEY WORDS: CANCER/TESTIS ANTIGENS; INTRINSICALLY DISORDERED PROTEINS; DOSAGE SENSITIVITY; CANCER

Intrinsically disordered proteins (IDPs) are proteins that, under physiological conditions in vitro, lack rigid 3D structures either along their entire length or in localized regions. Despite the lack of structure, the IDPs appear to play important biological roles in transcriptional regulation and signaling via cellular protein networks [Uversky and Dunker, 2010]. A comprehensive study of protein interaction networks in multiple eukaryotic organisms from yeast to human demonstrated that hub proteins, defined as those that interact with $\geq$5 partners in a protein interaction network, are significantly more disordered than end proteins, defined as those that interact with far fewer partners [Patil et al., 2010]. Furthermore, a binary classification of hubs and ends into ordered and disordered subclasses showed a significant enrichment of entirely disordered proteins and a significant depletion of entirely ordered proteins in hubs relative to ends [Haynes et al., 2006] underscoring the role of IDPs in signaling. Another interesting feature of the IDPs is their ability to undergo disorder-to-order transitions upon binding to their biological target (coupled folding and binding) in order to perform their function [Tompa and Csermely, 2004]. Structural flexibility and plasticity are believed to represent a major functional

Abbreviations used: CTAs, cancer/testis antigens; CT-X antigens, cancer/testis antigens on X chromosome; non-X CT antigens, cancer/testis antigens not on X chromosome; IDPs, intrinsically disordered proteins; RONN, regional order neural network; PRMT, protein arginine methyl transferase.

[†]These two authors made equal contributions and hence, they should both be considered as 2nd authors.

Additional Supporting Information may be found in the online version of this article.

Grant sponsor: NCI SPORE Grant; Grant number: 2P50CA058236-16; Grant sponsor: Patrick C Walsh Prostate Cancer Research Fund; Grant sponsor: PSOC Grant; Grant number: NCI U54 CA143803; Grant sponsor: NIDDK O'Brien Grant; Grant number: P50DK082998.

*Correspondence to: Prakash Kulkarni, Departments of Urology and Oncology, Johns Hopkins University School of Medicine, 600 N Wolfe St, Marburg 105B, Baltimore, MD 21287. E-mail: pkulkar4@jhmi.edu

Received 24 June 2011; Accepted 27 June 2011 • DOI 10.1002/jcb.23252 • © 2011 Wiley Periodicals, Inc.
Published online 11 July 2011 in Wiley Online Library (wileyonlinelibrary.com).

advantage for the IDPs enabling them to interact with a broad range of binding partners such as, proteins, nucleic acids, and small molecules [Tompa and Csermely, 2004].

Intrinsic disorder also appears to be an important determinant of dosage sensitivity. IDPs are prone to initiate promiscuous molecular interactions when overexpressed suggesting that this is the likely cause of the resulting toxicity/pathology. Indeed, recent studies in model organisms provide compelling evidence supporting this causality [Vavouri et al., 2009]. Interestingly, the same properties are strongly associated with dosage sensitive oncogenes, suggesting that mass action driven molecular interactions may be a frequent cause of cancer [Vavouri et al., 2009]. In fact, numerous IDPs are also associated with several other human diseases [Uversky et al., 2008] underscoring the tight association between intrinsic protein disorder, promiscuity, and dosage sensitivity.

The cancer/testis antigens (CTAs) are a heterogeneous group of proteins that are typically expressed in the testis with little or no expression in most somatic tissues. However, they are aberrantly expressed in several cancers [Scanlan et al., 2004], and recent genetic studies in the fruit fly have demonstrated a causal link between CTA expression and tumorigenesis [Janic et al., 2010]. Based on their chromosomal location, the CTAs can be conveniently divided into two broad groups: The CT-X antigens located on the X chromosome and non-X CT antigens located on the autosomes. Interestingly, most if not all, CT-X antigens lack orthologues in lower mammals and are found only in the primates where they constitute several subfamilies of homologous genes organized in discrete clusters along the X chromosome [Stevenson et al., 2007]. However, unlike the non-X CT antigens, the functions of a majority of the CT-X antigens are poorly understood although their overexpression is frequently associated with advanced disease and poorer prognosis [Suyama et al., 2010 and cf therein].

Given that intrinsic disorder is an important determinant of dosage sensitivity we asked if the CTAs, particularly the CT-X antigens, are IDPs as a result of their perceived pathological effects due to overexpression in advanced disease. Furthermore, our recent observations that the CT-X antigen, prostate-associated gene protein 4 (PAGE4), that is upregulated in prostate cancer is an IDP and that, its forced expression results in enhanced cell proliferation suggesting its dosage sensitive potential [Zeng et al., 2011], motivated us to undertake the present study.

MATERIALS AND METHODS

Disorder predictions in the CTAs were done applying the Foldindex [Prilusky et al., 2005] and regional order neural network (RONN) [Yang et al., 2005] algorithms and in some cases, metaPrDOS [Ishida and Kinoshita, 2008] was also employed in addition. To discern the effect of helical regions on protein disorder prediction, we compiled data using psiPred (http://bioinf.cs.ucl.ac.uk/psipred/) and JPred (http://www.compbio.dundee.ac.uk/www-jpred/) on a set of CT-X and non-X CTAs selected randomly before and after masking these regions. However, we found no difference in the prediction results presumably due to the paucity of helical regions in the disordered portions and therefore, we did not mask them in any of the analyses presented here. Based on the fraction of the sequence that was predicted to be disordered, we classified the CT-X and non-X CTAs into one of three classes: Highly ordered, (0–10% of the sequence is disordered), moderately disordered, (11–30% of the sequence is disordered), and highly disordered (31–100% of the sequence is disordered). To normalize for the varying protein lengths, we calculated the number of sequence motifs per 100 amino acids. PEST motifs were predicted using the epestfind algorithm of the EMBOSS package (http://emboss.bioinformatics.nl/cgi-bin/emboss/epestfind). Only motifs with a threshold PEST score >5 were considered. Ubiquitylation sites were predicted using UbPred [Radivojac et al., 2010]. Only the ubiquitylation sites with a high confidence score (range $0.84 \geq s \leq 1.00$) were considered. CTAs with percent ubiquitylation having a minimum value of 2, was used as a cutoff. Phosphorylation sites were predicted using KinasePhos 2.0 [Wong et al., 2007] which predicts the location of phosphorylation sites on S, T, and Y residues with a prediction specificity of 100%. CTAs with per cent phosphorylation having a minimum value of 2, was used as a cutoff. Acetylation sites were predicted using PAIL [Li et al., 2006] which predicts the acetylation sites on lysine residues with a high stringency and threshold score $\geq$0.5. Again, CTAs with percent acetylation having a minimum value of 3 was used as a cutoff. The probability to bind DNA was predicted using DBSPred [Ahmad et al., 2004] with a sensitivity setting of "strict." Arginine methylation sites were predicted using MEMO [Chen et al., 2006] and sumoylation sites were predicted using SUMOsp 2.0 [Ren et al., 2009]. Protein–protein interactions were predicted using the STRING interaction database [Jensen et al., 2009] at medium confidence setting (0.4–0.7) with no more than 10 interactions. The statistical analyses used to estimate significance were, Wilcoxon rank-sum, two sample *T*-test, and Chi square test as described in the text. The TATA box in the CTA promoter regions and specific sequence motifs in the mRNAs representing various polyadenylation and stability signals were searched by writing PERL scripts for each motif. Data in the CIRCOS plots were displayed by employing specific PERL scripts.

RESULTS

A CATALOG OF CTAs

As a first step in this direction, we constructed a comprehensive catalog of CTAs from the literature as well as the Cancer/Testis Antigen database (http://www.cta.lncc.br) [Almeida et al., 2009]. We identified 228 unique CTAs (Supplemental Table 1) and mapped them to their respective chromosomal locations. The CIRCOS plot in Figure 1 provides a detailed and comprehensive visual image of the location and density of the CTAs on each chromosome. Of the 228 CTAs, 120 CTAs (52%) mapped to the X chromosome (the CT-X antigens) while the remaining (non-X CT antigens), were distributed on the 22 autosomes and the Y chromosome. Among the autosomes, there are 10 CTAs (0.3 CTAs/100 genes) on chromosome 1 the most gene-rich chromosome. In contrast, chromosome 21 with only 425 genes has 1.6 CTAs/100 genes—a fivefold increase over chromosome 1—making it the most CTA-dense autosome while chromosome 7 with 0.06 CTAs/100 genes is the most CTA-poor chromosome. Among the sex chromosomes, while only 1 CTA is

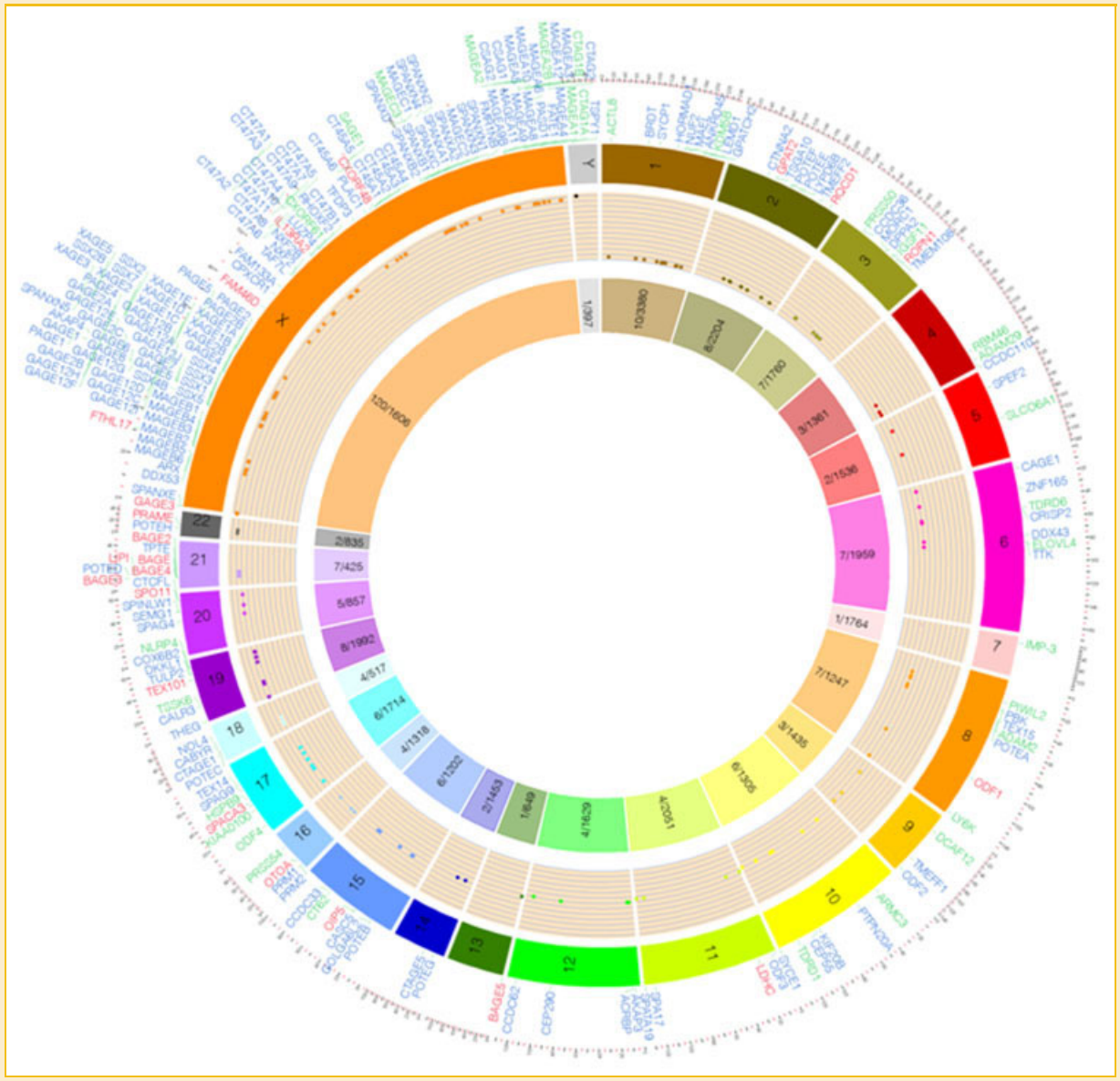

Fig. 1. CIRCOS plot showing the organization and disorder content of the cancer/testis antigens. The following information is presented going from the outside to the inside of the CIRCOS circles: Text Track—shows the names of all the CTAs. The highly ordered cancer/testis antigens (CTAs) (0–10% disorder) are indicated in red. The moderately disordered cancer/testis antigens (CTAs) (11–30% disorder) are indicated in green and highly disordered CTAs (31–100%) are shown in blue. Each track is drawn in order of its position on the respective chromosome. Scale Track—scale is reduced to 1e-6 and is shown in multiples of 10. Ideogram Track—colored track with numbers of the chromosomes. Scatter Plot—the CTAs are represented as solid circles based on their position on the chromosomes and colored to correspond with the chromosome they belong. The Track is divided into 13 lines and 12 spaces between them to show position of chromosomes 1–22, and the X and Y (innermost line indicates 0). Highlight Track—the transparent colored track showing the number of CT genes and the total number of genes on each chromosome.

present on the Y chromosome, there are 7.5 CTAs/100 genes on the X chromosome—a 25-fold increase when compared to chromosome 1 but a 125-fold increase over chromosome 7, the most CTA-poor chromosome (Fig. 1 and Supplemental Table 2).

A MAJORITY OF THE CTAs ARE INTRINSICALLY DISORDERED PROTEINS

We applied two different algorithms namely, FoldIndex [Prilusky et al., 2005] and RONN [Yang et al., 2005], to predict protein disorder. FoldIndex implements an algorithm to make a calculation based on average net charge and average hydrophobicity of the sequence to predict whether a given sequence is ordered or disordered. In contrast, RONN uses a neural network technique to predict whether any given residue is likely to be ordered or disordered in the context of the surrounding amino acid sequence. Although the physical properties of amino acids are the fundamental basis in determining disorder, the neural network used in RONN avoids explicit parameterization of amino acids in such a manner. Instead it uses non-gapped sequence alignment to measure "distances" between windows of sequence for the unknown protein

and windowed sequences for known folded proteins derived from the Protein Database (PDB). Therefore, FoldIndex and RONN represent two fundamentally different approaches to disorder prediction and while both methods have their strengths and weaknesses, they perform well when compared to other disorder prediction methods. Nonetheless, FoldIndex performs particularly well for fully ordered or fully disordered sequences, while RONN is more successful in identifying partially disordered sequences. Thus, employing these two prediction models, we classified the CTAs either separately or collectively into three groups based on the extent of disorder: highly ordered, moderately disordered, and highly disordered (see Materials and Methods Section).

As shown in Figure 2A, a vast majority of the CTAs (>90%) belong to the intrinsically disordered class of proteins regardless of the prediction method ($\chi^2 = \text{NS}$). When examined separately, both prediction methods (Foldindex, Fig. 2B, and RONN, Fig. 2C) demonstrated that the CT-X rather than the non-X CT antigens were significantly more disordered (χ^2: $P < 0.0001$). However, in either case, the majority of both the CT-X and non-X CT antigens were in the highly disordered group. The details of the disorder predictions for the CTAs in each group both by FoldIndex and RONN are presented in Supplemental Tables 3–6, respectively.

Despite the strong agreement between the prediction methods in the vast majority of the cases we observed some differences either in the extent of disorder or the regions of disorder in a few instances. In such cases we used an additional prediction method namely, metaPrDOS. metaPrDOS which uses a meta approach, does not predict disordered regions from amino acid sequence directly but predicts them by integrating the results of eight distinct prediction methods [Ishida and Kinoshita, 2008]. However, the results predicted by metaPrDOS were similar to those predicted separately by Foldindex or RONN and therefore, we applied Foldindex to predict disorder in all subsequent analyses presented here. However, data were also obtained by subjecting the CTAs to similar analyses by RONN and are presented in the Supplemental Online Material.

REGULATION OF INTRACELLULAR CTA CONCENTRATIONS

Given that the altered abundance of several IDPs is associated with perturbed cellular signaling that may lead to pathological conditions such as cancer, it is important to understand how the cellular concentrations of IDPs are precisely regulated. Indeed, recent studies on the yeast and human proteome have revealed that there is an evolutionarily conserved tight control of synthesis and clearance of most IDPs [Gsponer et al., 2008; Edwards et al., 2009]. We therefore examined the CTAs at the genomic, transcript, and protein levels to discern how their intracellular concentrations may be regulated and whether the regulation correlates with protein disorder.

CTA CONCENTRATIONS MAY NOT BE REGULATED AT THE TRANSCRIPT LEVEL

We first examined genomic sequences encoding the CTAs for the presence of a TATA box in the promoter region. However, a preliminary analysis suggested that contrary to previous observations [Gsponer et al., 2008], there did not appear to be a correlation between the presence/absence of the TATA box and protein disorder and hence, we did not undertake a detailed analysis. Next we looked

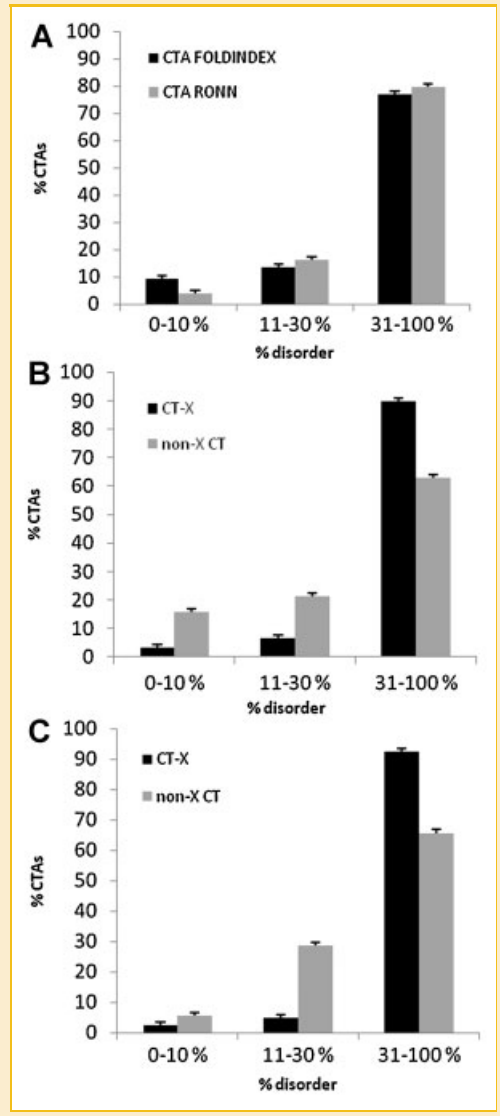

Fig. 2. Predicting disorder in the cancer/testis antigens. Protein disorder was determined in all the cancer/testis antigens (CTAs) using both Foldindex and RONN (A). The CTAs were divided into 3 groups based on the extent of disorder: highly ordered (0–10% disorder), moderately disordered (11–30% disorder), and highly disordered (31–100% disorder). Protein disorder prediction was also done separately on the CT-X and non-X groups using Foldindex (B) and RONN (C). Standard errors were calculated and all reported differences were found to be statistically significant (Chi square test: NS for A, $P < 0.0001$ for B and C). NS = not significant.

at the transcript level and examined the sequences associated with mRNA stability/turnover. For the presence of polyadenylation signals (PASs), we searched for the following motifs that have been reported in the literature: 5′AGUAAA3′ (PAS 1); 5′AAUAAA/AUUAAA/AAUAAA3′ (PAS 2); 5′UAUAAA3′ (PAS 3); 5′CAUAAA3′ (PAS 4); 5′GAUAAA3′(PAS 5) [Beaudoing et al., 2000]. For RNA stability we searched for multiple signals including, PUM-binding sites (5′UGUACAUA/UAUA/AAUA3′) [Galgano et al., 2008], U-rich motif(s) (URM) (5′UUUUAAA/UUUGUUU3′) [Bolognani et al., 2010], the stability sequence (5′UAUUUAU3′) [Wiklund et al., 2002], cytoplasmic polyadenylation element (CPE) (5′UUUUUAU3′) [Mor-

gan et al., 2010], and the heptanucleotide AU-rich element ARE motif (5′UAUUUAU3′) [Barreau et al., 2005], both in the entire transcript as well as in only the 3′ untranslated regions. Human PUM1 and PUM2 are members of the Puf family an evolutionarily conserved family of RNA-binding proteins related to the Pumilio proteins of *Drosophila* and the fem-3 mRNA binding factor proteins of *C. elegans*. The encoded proteins contain a sequence-specific RNA binding domain and serve as translational regulators of specific mRNAs by binding to their 3′ untranslated regions [Spassov and Jurecic, 2002]. However, in contrast to previous observations [Gsponer et al., 2008], we did not observe any significant correlation between the presence/absence of these motifs in the mRNA and the extent of disorder in the CTAs encoded by them (Supplemental Tables 7–46).

CTA CONCENTRATIONS MAY BE REGULATED AT THE PROTEIN LEVEL

Next, we examined the CTA protein sequences to discern sequence motifs characteristic of protein turnover/stability. The PEST sequence is thought to be a hallmark of protein degradation and stability [Rechsteiner and Rogers, 1996]. Employing the epestfind algorithm (http://emboss.bioinformatics.nl/cgi-bin/emboss/epestfind), we observed a significant increase in the number of PEST motifs that was directly proportional to the amount of disorder (χ^2: $P < 0.0006$) (Fig. 3A). Separating the CTAs into CT-X and non-X also showed a similar trend; there was a significant correlation between the number of PEST motifs and extent of disorder (Wilcoxon Rank Sum Test: $P = 0.0231$ and 0.0164, respectively) (Fig. 3B and C). Overall, however, the CT-X antigens appeared to have a significantly higher fraction of proteins with the PEST motif than did the non-X CTAs (*T*-test: $P < 0.02$) (Fig. 3D). We also performed similar analyses on the CTAs correlated with protein disorder predicted using RONN. Again, the results were comparable to those obtained with Foldindex (Supplemental Fig. 2A–D). The details of the PEST analyses by both disorder prediction methods are presented in Supplemental Tables 47–50.

Ubiquitylation is another covalent protein modification that is frequently associated with proteasome-mediated degradation [Welchman et al., 2005]. By employing the UbPred algorithm [Radivojac et al., 2010], we observed a significant correlation between the occurrence of the consensus ubiquitylation site and CTA disorder content (χ^2: $P = 0.001$) (Fig. 4A). When analyzed separately, both in CT-X and non-X CTAs, there was a significant association between the presence of the ubiquitylation site and extent of disorder (Wilcoxon Rank Sum Test: $P < 0.0001$) (Fig. 4B and C). However, there was no difference between the

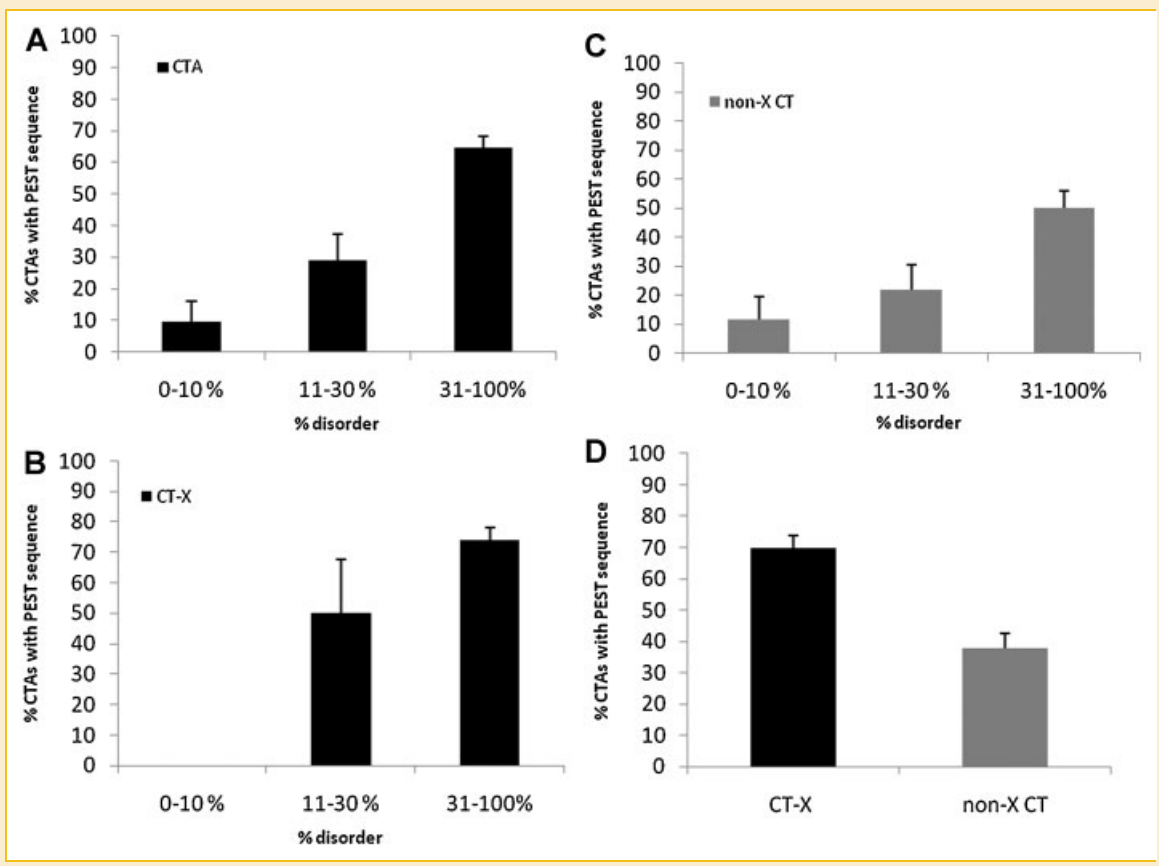

Fig. 3. Correlation between presence of PEST sequence and extent of disorder in the cancer/testis antigens. Percent cancer/testis antigens (CTAs) with PEST sequence/100 amino acids (A). Percent CTAs with PEST sequences seen in the three disordered groups of CT-X (B), and non-X CT antigens (C), respectively. CTAs were segregated into CT-X and non-X CT antigens and percent CTAs with PEST sequences were plotted (D). The Foldindex algorithm was applied to group the CTAs. Standard errors were calculated and all reported differences were found to be statistically significant (Chi square test: $P < 0.001$ for A, Wilcoxon Rank Sum Test (RS): $P < 0.05$ for B and C and *T*-test: $P < 0.05$ for D).

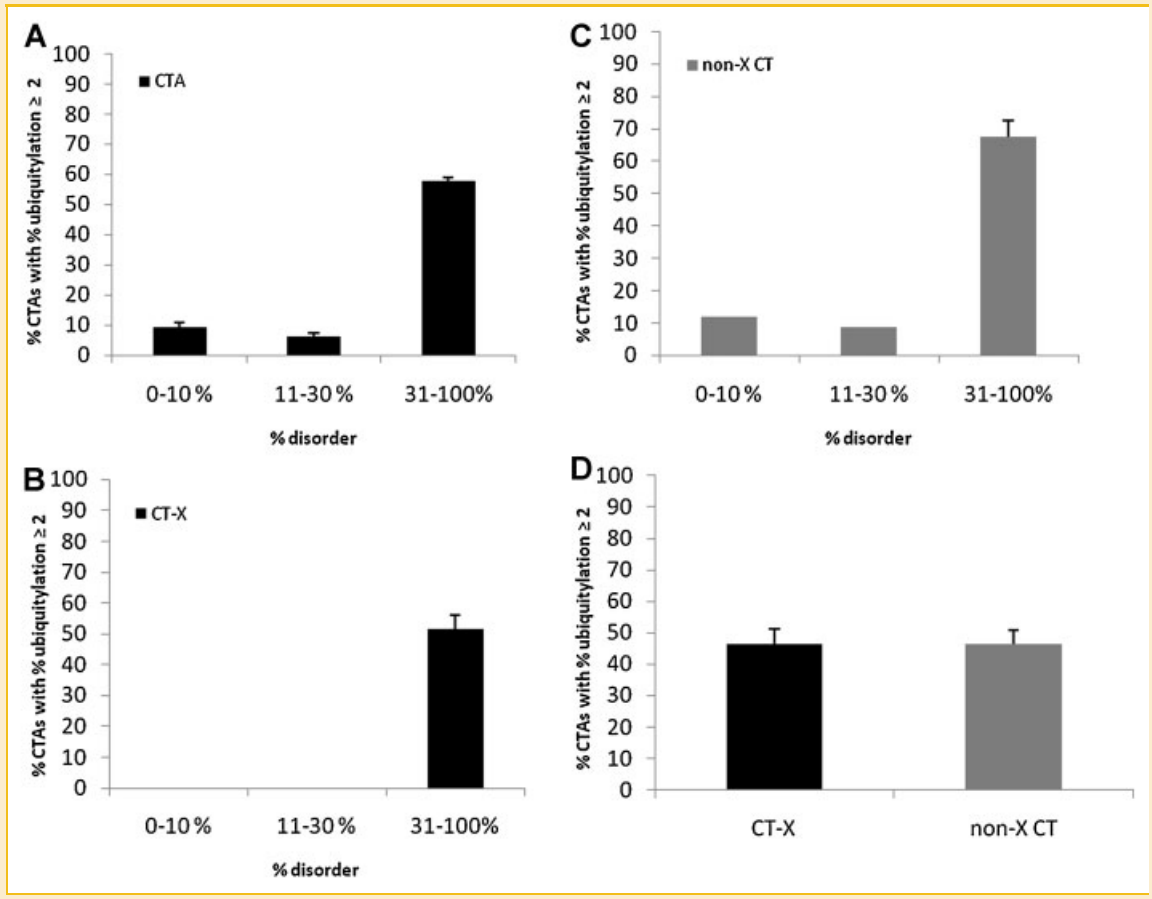

Fig. 4. Correlation between presence of ubiquitylation sites and disorder in the cancer/testis antigens. The percent CTAs with ≥2 ubiquitylation sites/100 amino acids are plotted as a function of disorder calculated by Foldindex (A). CT-X and non-X CT antigens were then plotted separately with respect to disorder (B and C). CTAs were segregated into CT-X and non-X CT antigens and percent CTAs with percent ubiquitylation sites ≥2 were plotted (D). Standard errors were calculated and all reported differences were found to be statistically significant (Chi square test: $P < 0.001$ for A, Wilcoxon Rank Sum Test (RS): $P < 0.001$ for B and C, and *T*-test: not significant for D).

two groups, CT-X and non-X CTAs, when considered in the absence of disorder content (*T*-test: NS) (Fig. 4D). We also performed similar analyses on the CTAs correlated with protein disorder predicted using RONN. Again, the results were comparable to those obtained with Foldindex (Supplemental Fig. 3A–D). The details of the ubiquitylation analyses by both disorder prediction methods are presented in Supplemental Tables 51–54. Considered together, these data on the messenger RNA and protein turnover/stability suggested that unlike most IDPs [Gsponer et al., 2008; Edwards et al., 2009], the CTAs do not appear to be regulated at the mRNA synthesis or stability level but instead, appear to be regulated at the protein level.

DISORDERED CTAs ARE SIGNIFICANTLY MORE LIKELY TO BE MODIFIED BY PHOSPHORYLATION AND ACETYLATION

Covalent modification by phosphorylation is also frequently observed in IDPs [Iakoucheva et al., 2004] and is thought to play a critical role in their functions [Galea et al., 2008]. Thus, employing the KinasePhos 2.0 algorithm [Wong et al., 2007] that predicts phosphorylation at S, T, and Y residues we examined the CTAs for the presence of the respective consensus motifs. As shown in (Fig. 5A), although there was no difference between the highly ordered and moderately disordered CTAs, the highly disordered CTAs were significantly enriched for these motifs (χ^2: $P = 0.0044$). In both groups, the highly disordered CTAs were significantly enriched for these motifs (Wilcoxon Rank Sum Test: $P = 0.0166$ and 3×10^{-6}, respectively) (Fig. 5B and C). Between the two groups however, the CT-X antigens appeared to have significantly more phosphorylation sites than the non-X CT (*T*-test: $P < 0.02$) (Fig. 5D). We also performed similar analyses on the CTAs correlated with protein disorder predicted using RONN. Again, the results were similar to those obtained with Foldindex (Supplemental Fig. 4A–D). The details of the phosphorylation analyses by both disorder prediction methods are presented in Supplemental Tables 55–58.

Protein acetylation at lysine residues that plays an important role in various biological processes [Arif et al., 2010], also appears to be important in modulating the functions of many IDPs [Hansen, 2006; van Dieck et al., 2009]. Thus, we examined the CTAs for potential lysine acetylation employing the PAIL algorithm [Li et al., 2006]. As shown in (Fig. 6A), we observed a significant correlation between CTA protein disorder and the presence of acetylated lysines (χ^2: $P = 0.0067$). In both groups, the highly disordered CTAs were significantly enriched for these motifs (Wilcoxon Rank Sum Test: $P = 0.0055$ and $P < 0.0001$, respectively) (Fig. 6B and C). Between the two groups however, the CT-X antigens appeared to have

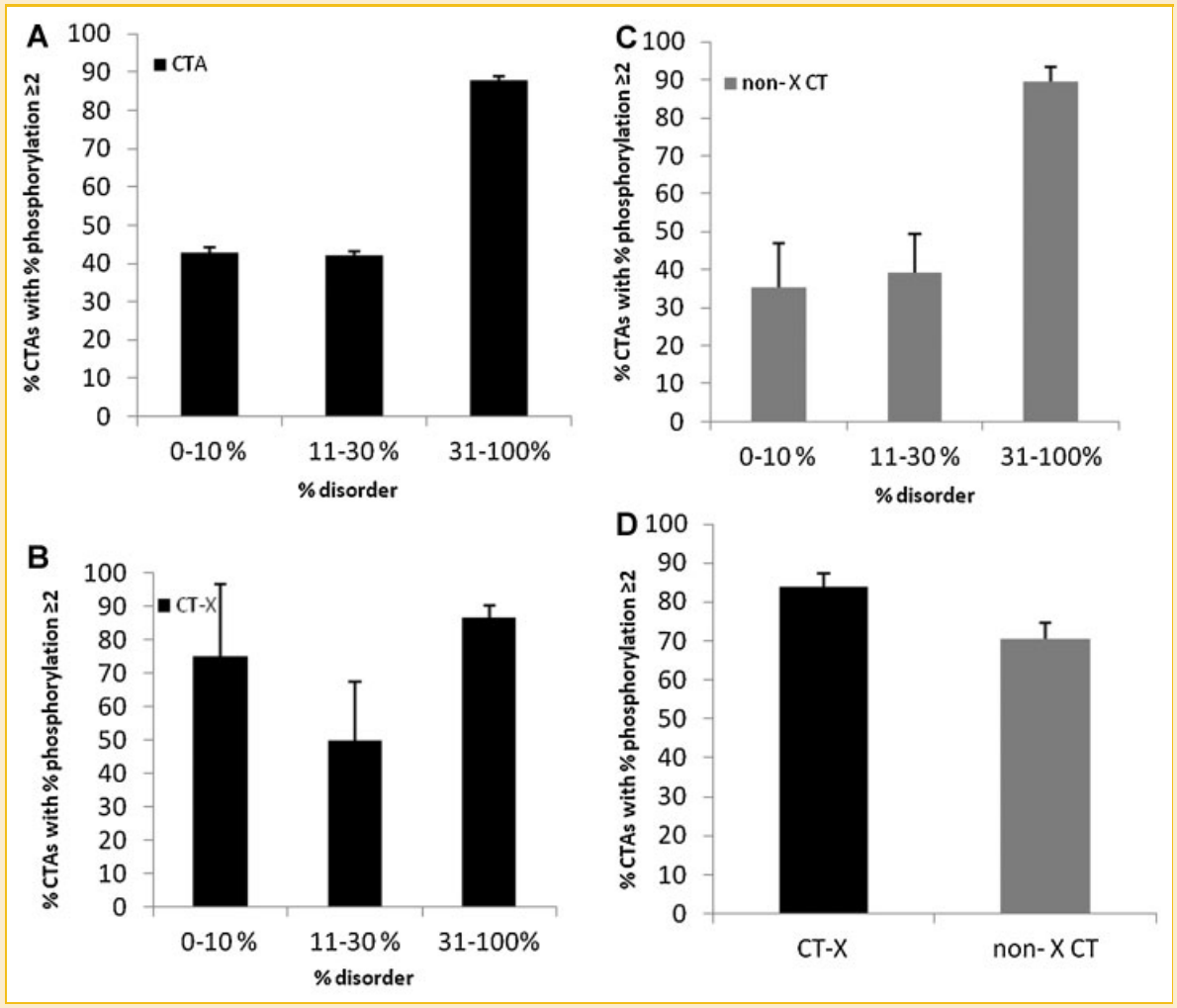

Fig. 5. Correlation between the presence of phosphorylation sites and disorder in the cancer/testis antigens. Percent cancer/testis antigens (CTAs) with ≥2 phosphorylation sites/100 amino acids is plotted with respect to disorder (A). Percent CT-X (B) and non-X CT antigens (C) with ≥2 phosphorylation sites/100 amino acids was plotted with respect to disorder, respectively. Phosphorylation sites were predicted for both CT-X and non-X CT antigens (D). The Foldindex algorithm was applied to group the CTAs. Standard errors were calculated and all reported differences were found to be statistically significant (Chi square test: $P < 0.01$ for A, Wilcoxon Rank Sum Test (RS): $P < 0.05$, $P < 0.001$ for B and C, respectively, and *T*-test: $P < 0.05$ for D).

significantly less acetylation sites than the non-X CT antigens (*T*-test: $P < 0.03$) (Fig. 6D). The details of the acetylation analyses by both disorder prediction methods are presented in Supplemental Tables 59–62. The results obtained using RONN are shown in Supplemental Figure 5A–D. Considered together, covalent modifications of the CTAs by phosphorylation and acetylation may play critical roles in modulating their interactions by altering the local physicochemical properties of the intrinsically disordered CT proteins/regions.

CTAs LACK MODIFICATIONS BY ARGININE METHYLATION AND SUMOYLATION

Protein methylation particularly, lysine methylation, is frequently observed in many organisms. Thus, major attention has been focused on lysine methylation because of its role in chromatin remodeling and transcriptional regulation, emerging evidence suggests that arginine methylation may also play an important role in many physiological processes such as signal transduction, mRNA splicing, transcriptional control, DNA repair, and protein translocation [Bedford and Clarke, 2009]. Furthermore, since the covalent marking of proteins by arginine methylation can promote their recognition by binding partners or can modulate their biological activity it was of interest to interrogate the CTAs, many of which are implicated in similar functions, for arginine methylation. To this end, we applied the algorithm MEMO [Chen et al., 2006] that identifies specific arginine residues that are likely to get methylated by protein arginine methyl transferase (PRMT). Indeed, the program predicted several arginine residues as highly likely to be methylated by PRMT. However, we did not observe any significant difference in the extent of arginine methylation and protein disorder (Supplemental Tables 63–66).

SUMOylation is a post-translational modification that is involved in various cellular processes, such as cell cycle regulation, gene transcription, differentiation, cellular localization apoptosis, protein

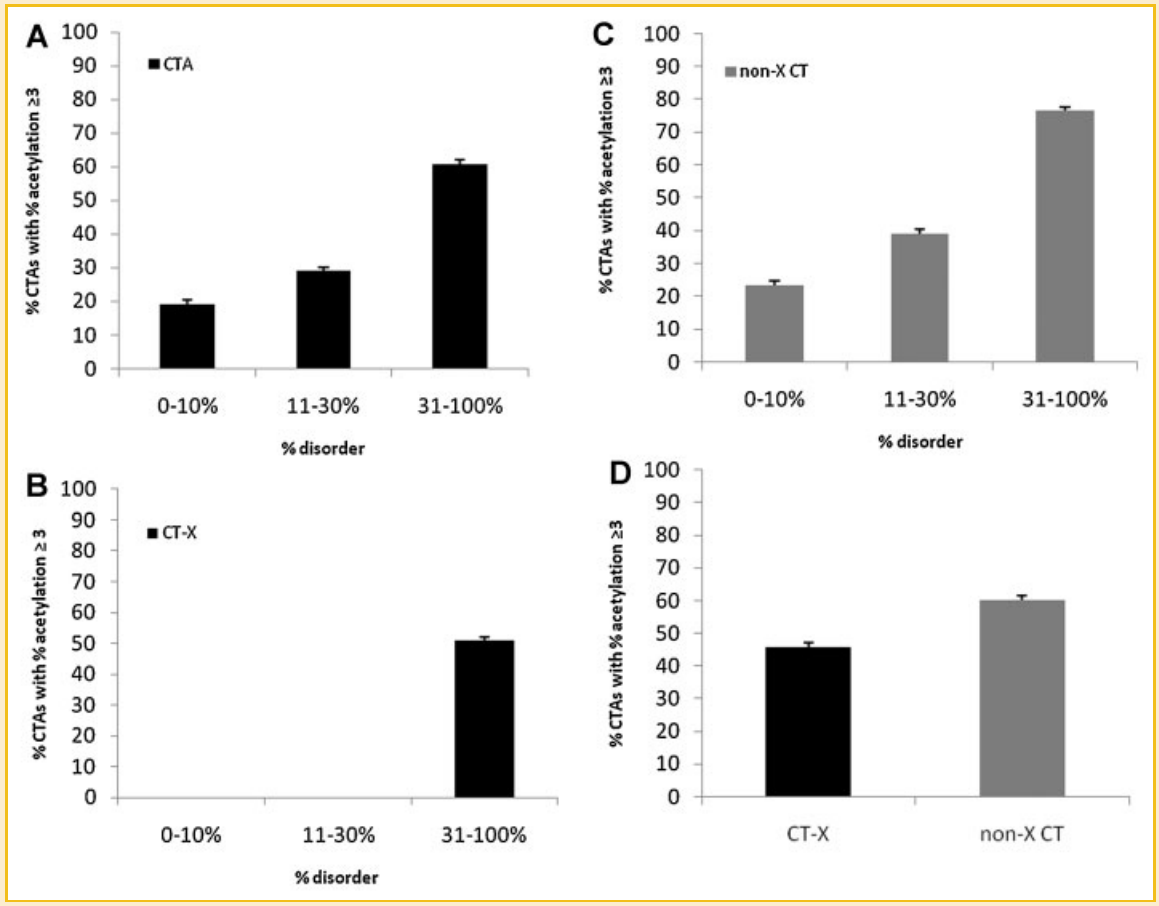

Fig. 6. Correlation between presence of acetylation sites and disorder in the cancer/testis antigens. The percent cancer/testis antigens (CTAs) with ≥3 acetylation sites/100 amino acids is plotted with respect to disorder (A). Percent CT-X (B) and non-X CT antigens (C) with ≥3 acetylation sites/100 amino acids was plotted with respect to disorder. Acetylation sites were predicted for both CT-X and non-X CT antigens (D). The Foldindex algorithm was applied to group the CTAs. Standard errors were calculated and all reported differences were found to be statistically significant (Chi square test: $P < 0.01$ for A, Wilcoxon Rank Sum Test (RS): $P < 0.05$, $P < 0.001$ for B and C, respectively, and *T*-test: $P < 0.05$ for D).

stability, response to stress, and progression through the cell cycle [Hannoun et al., 2010; Mooney et al., 2010]. Since a majority of CTAs are IDPs and therefore participate in many of these physiological processes, we asked if there is a correlation between CTA disorder and SUMOylation employing the SUMOsp2.0 algorithm [Ren et al., 2009]. However, we did not observe any significant correlation (Supplemental Tables 67–70).

CORRELATION BETWEEN DNA-BINDING PROBABILITY AND PROTEIN DISORDER

Recently, Liu et al. [2009] presented a quantitative theory predicting the role of intrinsic disorder in protein structure and function by applying thermodynamic models of protein interactions in which IDPs are characterized by positive folding free energies. The authors used the Gene Ontology classifications "protein binding," "catalytic activity," and "transcription regulator activity," and performed genome-wide surveys of both the amount of disorder and the binding affinities in these functional classes for prokaryotic and eukaryotic genomes. Specifically, without assuming any a priori structure-function relationship, their theory predicted that both catalytic and low-affinity binding ($K_d \geq 10^{-7}$ M) proteins prefer ordered structures, whereas only high-affinity binding proteins (found mostly in eukaryotes) can tolerate disorder. Furthermore, of particular relevance to both transcription and signal transduction, the theory also explained how increasing disorder can tune the binding affinity to maximize the specificity of promiscuous interactions [Liu et al., 2009].

Thus, we asked if the CTAs may also be associated with transcriptional regulation and hence, bind DNA in light of their disordered structure. To this end we employed the DBS-Pred algorithm [Ahmad et al., 2004] to predict the probability of DNA binding at two different stringencies. As shown in (Fig. 7A), using a cutoff of 50%, we observed a significant correlation between DNA binding prediction probability and extent of protein disorder (χ^2: $P = 0.0001$). Similar results were obtained when the CTAs were divided into the CT-X and non-X groups (Fig. 7B and C, respectively). Increasing the stringency (>90% prediction probability) also yielded similar results; a majority of the highly disordered CTAs were predicted to bind DNA with virtually none in the moderately disordered group (Fig. 7D). As expected, there were significantly fewer CTAs with DNA binding probability when the stringency was increased. However, the drop was more pronounced

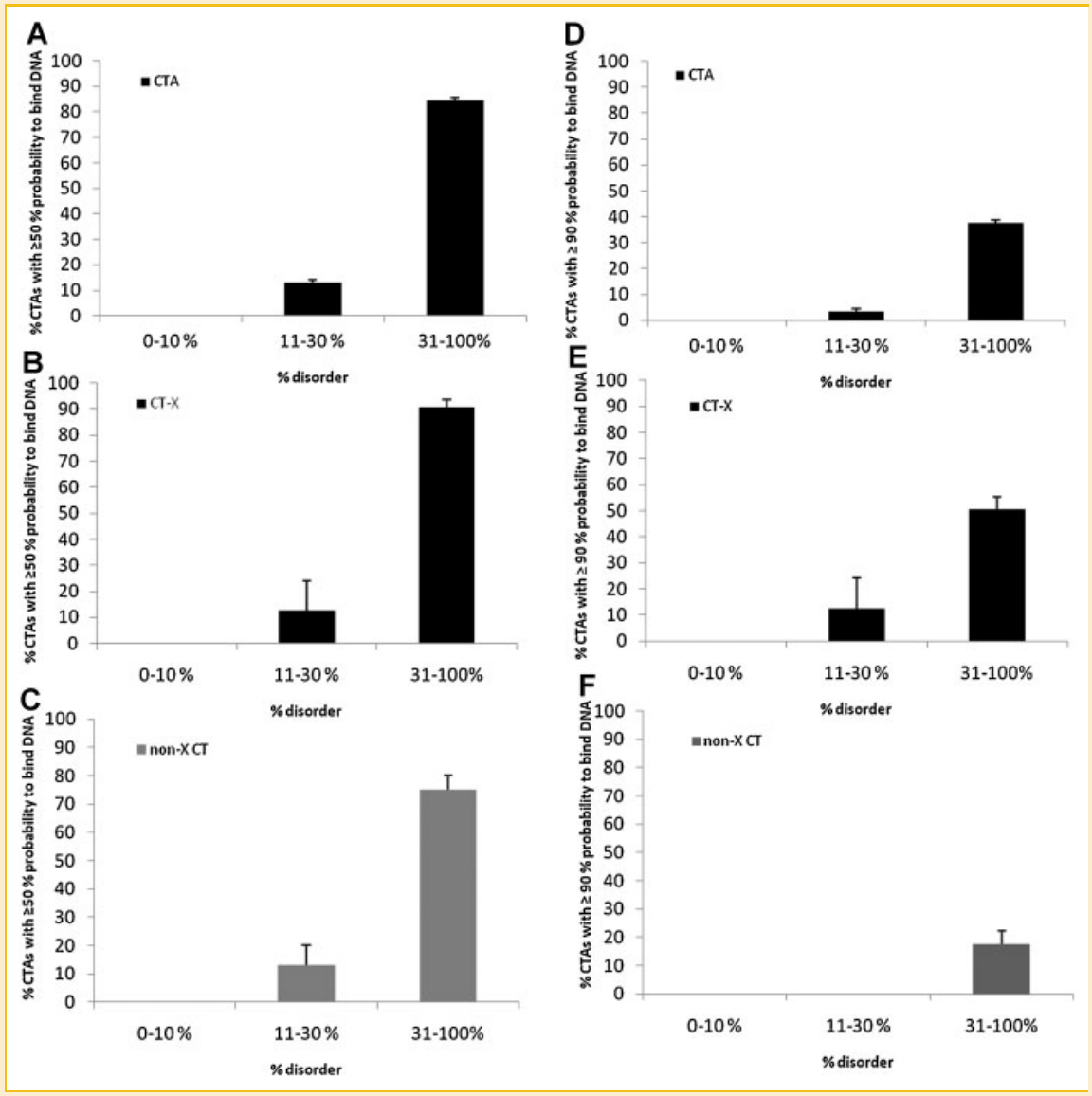

Fig. 7. Correlation between DNA binding prediction probability and disorder in the cancer/testis antigens. Percent cancer/testis antigens (CTAs) with probability of binding DNA was determined applying DBS-PRED with ≥50% probability (A–C) or ≥90% probability (D–F) of binding DNA without grouping (A and D) or with grouping the CTAs into CT-X (B and E) and non-X CTAs (C and F), respectively. The Foldindex algorithm was applied to group the CTAs. Standard errors were calculated and all reported differences were found to be statistically significant (Chi square test: $P < 0.001$ for A and D, Wilcoxon Rank Sum Test (RS): $P < 0.001$ for B, C, E, and F).

in the non-X group than in the CT-X group (Fig. 7F and E, respectively). Consistent with their potential DNA-binding function, several studies have demonstrated that many of the CT-X antigens are localized in the nucleus [Westbrook et al., 2004; Bai et al., 2005; Zhao et al., 2011]. Taken together, these data suggested that indeed, the CTAs are likely to be involved in transcriptional regulation or other processes such as DNA damage/repair or chromatin remodeling, for example, that involve DNA. The details of the DBSPred analyses are presented in Supplemental Tables 71–74.

CTAs OCCUPY HUB POSITIONS IN PROTEIN–PROTEIN INTERACTION NETWORKS

As mentioned earlier, IDPs typically occupy hub positions in a protein interaction network [Patil et al., 2010]. To determine if indeed this was also the case with the CTAs that we predicted to be disordered, we selected representative members from the CT-X antigens with as yet unknown functions. We determined their putative interactions by querying STRING, a database dedicated to protein–protein interactions that include both physical and functional interactions through the so-called "genomic context" or "nonhomology-based" inference methods [Jensen et al., 2009]. As shown in Figure 8A–F, most CT-X antigens occupy a hub position. Consistent with the propensity of IDPs to function in transcriptional regulation and/or cellular signaling, the data suggest, but do not necessarily prove, that the CT-X antigens may also fulfill such roles in the cell. Furthermore, many of the CTAs in these networks have previously been demonstrated to participate in transcriptional regulation making our conclusion more tenable.

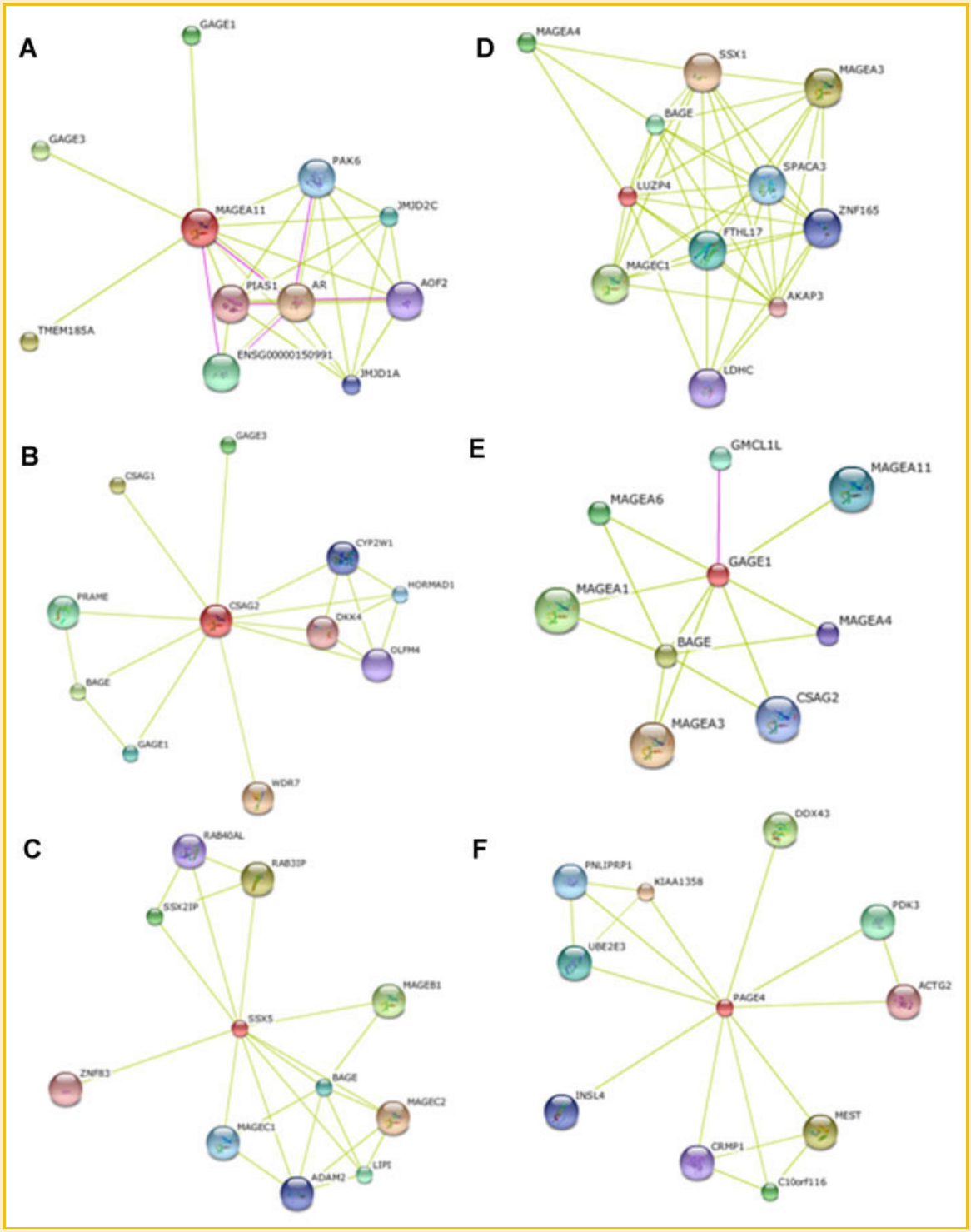

Fig. 8. Protein–protein interactions involving CT-X antigens. Protein–protein interactions were derived by querying the STRING database. CT-X antigens with disorder content ranging from 50% to 100% were randomly selected. As expected, each of the input CT-X antigens occupied hub positions in the network. In five of the six cases, the input CT-X antigens preferentially interact with other CTAs. (A) MAGEA11 = 49.6% disorder and 30.6% DNA-binding probability. (B) CSAG2 = 52.7% disorder and 58.5% DNA-binding probability. (C) SSX5 = 85% disorder and 98.6% DNA-binding probability. (D) LUZP4 = 91% disorder and 97.4% DNA-binding probability. (E) GAGE1 = 100% disorder and 96.5% DNA-binding probability. (F) PAGE4 = 100% disorder and 99.1% DNA-binding probability.

CTAs AND DOSAGE SENSITIVITY

To discern a potential causal link between aberrant CTA expression (increase in concentration) and dosage sensitivity (defined here as increased cell growth phenotype), we examined data from the literature. We compiled data on 41 experiments reporting either siRNA-mediated silencing or overexpression of specific CTAs. As expected, silencing gene expression in cells overexpressing specific CTAs resulted in decreased cell growth, while their forced expression in cells lacking expression, increased growth of the transfected cells (Supplemental Table 75). Together, these independent experiments on a variety of CT-X and non-X CT antigens provide good evidence supporting causality between CTA overexpression and dosage sensitivity in cancer.

DISCUSSION

Many proteins in living cells appear to be involved in the transfer and processing of information. Such proteins are functionally linked via networks to form biochemical ''circuits'' that perform a variety of simple computational tasks including information amplification, integration, and storage [Bray, 1995]. Emerging evidence applying network theory suggests that the architecture of such networks is not random but instead is ''scale-free'' with most proteins representing nodes having only a few connections and a relatively fewer proteins occupying ''hubs'' with tens, hundreds, or more links [Dunker et al., 2005; Almaas et al., 2007]. Scale-free networks are highly dynamic and grow incrementally. Interestingly, when ''deciding'' where to establish a link, a new node ''prefers'' an existing node that already has many connections (hub) over one with fewer links. These two basic mechanisms, growth and preferential attachment, will eventually lead to the system being dominated by hubs.

But what structural and functional attributes of a protein makes it ''desirable'' for recruitment to a hub position so that it can interact with a large number of diverse targets? A resounding answer appears to be an IDP because of the unique thermodynamic advantage IDPs posses by existing as an ensemble of very different conformations in fast exchange [Uversky, 2002], and their capability to adapt to new demands. Furthermore, because they are typically dosage sensitive, IDPs are more likely to participate in a large number of promiscuous interactions when overexpressed, simply as a consequence of mass action [Marcotte and Tsechansky, 2009; Vavouri et al., 2009].

A hallmark of such inhomogeneous scale-free networks is their resilience. Thus, for example, in yeast, although proteins with five or fewer links constitute about 93% of the total number of proteins, only about 21% of them are essential. In contrast, only about 0.7% of the yeast proteins with known phenotypic profiles have more than 15 links, but single deletion of as many as 62% proves lethal implying that highly connected proteins with a central role in the network's architecture are three times more likely to be essential than proteins with only a small number of links to other proteins [Jeong et al., 2001]. A take home lesson from these observations in yeast is that similar scale-free networks maybe operational in cancer making the disease so resilient. Perhaps our failure to combat the disease in spite of decades of intense research and 40 years of declaring ''war'' against cancer maybe due to that fact that we are targeting common nodes rather than the critical hubs.

Taken together, our data suggest that the CTAs by occupying hub positions in protein networks could create new nodes with novel functions leading to the observed pathological phenotype in the *absence* of genetic changes. Further, the data provide a novel perspective on the CTAs implicating them in processing and transducing information in altered physiological states in a dosage sensitive manner. Identifying CTAs that occupy hub positions in protein regulatory networks would allow a better understanding of their functions as well as the development of novel therapeutics to treat cancer.

ACKNOWLEDGMENTS

The authors would like to thank Dr. Amita Behal for her help with the bioinformatics analyses. SMM is supported by an American Urological Association Foundation Research Scholarship. This work was supported by NCI SPORE Grant 2P50CA058236-16, the Patrick C Walsh Prostate Cancer Research Fund (PK), a NIDDK O'Brien Grant P50DK082998, and PSOC Grant NCI U54 CA143803 (RGH).

REFERENCES

Ahmad S, Gromiha MM, Sarai A. 2004. Analysis and prediction of DNA-binding proteins and their binding residues based on composition, sequence and structural information. Bioinformatics 20:477–486.

Almaas E, Vazquez A, Barabasi AL. 2007. Complex systems and interdisciplinary science. Vol. 3, Singapore: World Scientific publishing Co. Pte. Ltd. pp 1–20.

Almeida LG, Sakabe NJ, deOliveira AR, Silva MC, Mundstein AS, Cohen T, Chen YT, Chua R, Gurung S, Gnjatic S, Jungbluth AA, Caballero OL, Bairoch A, Kiesler E, White SL, Simpson AJ, Old LJ, Camargo AA, Vasconcelos AT. 2009. CTdatabase: A knowledge-base of high-throughput and curated data on cancer-testis antigens. Nucleic Acids Res 37:D816–D819.

Arif M, Senapati P, Shandilya J, Kundu TK. 2010. Protein lysine acetylation in cellular function and its role in cancer manifestation. Biochim Biophys Acta 1799:702–716.

Bai S, He B, Wilson EM. 2005. Melanoma antigen gene protein MAGE-11 regulates androgen receptor function by modulating the interdomain interaction. Mol Cell Biol 25:1238–1257.

Barreau C, Paillard L, Osborne HB. 2005. AU-rich elements and associated factors: Are there unifying principles? Nucleic Acids Res 33:7138–7150.

Beaudoing E, Freier S, Wyatt JR, Claverie JM, Gautheret D. 2000. Patterns of variant polyadenylation signal usage in human genes. Genome Res 10:1001–1010.

Bedford MT, Clarke SG. 2009. Protein arginine methylation in mammals: Who, what, and why. Mol Cell 33:1–13.

Bolognani F, Contente-Cuomo T, Perrone-Bizzozero NI. 2010. Novel recognition motifs and biological functions of the RNA-binding protein HuD revealed by genome-wide identification of its targets. Nucleic Acids Res 38:117–130.

Bray D. 1995. Protein molecules as computational elements in living cells. Nature 376:307–312.

Chen H, Xue Y, Huang N, Yao X, Sun Z. 2006. MeMo: A web tool for prediction of protein methylation modifications. Nucleic Acids Res 34: W249–W253.

Dunker AK, Cortese MS, Romero P, Iakoucheva LM, Uversky VN. 2005. Flexible nets. The roles of intrinsic disorder in protein interaction networks. FEBS J 272:5129–5148.

Edwards YJ, Lobley AE, Pentony MM, Jones DT. 2009. Insights into the regulation of intrinsically disordered proteins in the human proteome by analyzing sequence and gene expression data. Genome Biol 10:R50.1–R50.18.

Galea CA, Nourse A, Wang Y, Sivakolundu SG, Heller WT, Kriwacki RW. 2008. Role of intrinsic flexibility in signal transduction mediated by the cell cycle regulator, p27 Kip1. J Mol Biol 376:827–838.

Galgano A, Forrer M, Jaskiewicz L, Kanitz A, Zavolan M, Gerber AP. 2008. Comparative analysis of mRNA targets for human PUF-family proteins suggests extensive interaction with the miRNA regulatory system. PLoS One 3(9):e3164.1–16.

Gsponer J, Futschik ME, Teichmann SA, Babu MM. 2008. Tight regulation of unstructured proteins: From transcript synthesis to protein degradation. Science 322:1365–1368.

Hannoun Z, Greenhough S, Jaffray E, Hay RT, Hay DC. 2010. Post-translational modification by SUMO. Toxicology 278:288–293.

Hansen JC. 2006. Linking genome structure and function through specific histone acetylation. ACS Chem Biol 1:69–72.

Haynes C, Oldfield CJ, Ji F, Klitgord N, Cusick ME, Radivojac P, Uversky VN, Vidal M, Iakoucheva LM. 2006. Intrinsic disorder is a common feature of hub proteins from four eukaryotic interactomes. PLoS Comput Biol 2(8):e100.0001–0012.

Iakoucheva LM, Radivojac P, Brown CJ, O'Connor TR, Sikes JG, Obradovic Z, Dunker AK. 2004. The importance of intrinsic disorder for protein phosphorylation. Nucleic Acids Res 32:1037–1049.

Ishida T, Kinoshita K. 2008. Prediction of disordered regions in proteins based on the meta approach. Bioinformatics 24:1344–1348.

Janic A, Mendizabal L, Llamazares S, Rossell D, Gonzalez C. 2010. Ectopic expression of germline genes drives malignant brain tumor growth in *Drosophila*. Science 330:1824–1827.

Jensen LJ, Kuhn M, Stark M, Chaffron S, Creevey C, Muller J, Doerks T, Julien P, Roth A, Simonovic M, Bork P, von Mering C. 2009. STRING 8—A global view on proteins and their functional interactions in 630 organisms. Nucleic Acids Res 37:D412–D416.

Jeong H, Mason SP, Barabasi AL, Oltvai ZN. 2001. Lethality and centrality in protein networks. Nature 411:41–42.

Li A, Xue Y, Jin C, Wang M, Yao X. 2006. Prediction of Nepsilon-acetylation on internal lysines implemented in Bayesian discriminant method. Biochem Biophys Res Commun 350:818–824.

Liu J, Faeder JR, Camacho CJ. 2009. Toward a quantitative theory of intrinsically disordered proteins and their function. Proc Natl Acad Sci USA 106:19819–19823.

Marcotte EM, Tsechansky M. 2009. Disorder, promiscuity, and toxic partnerships. Cell 138:16–18.

Mooney SM, Grande JP, Salisbury JL, Janknecht R. 2010. Sumoylation of p68 and p72 RNA helicases affects protein stability and transactivation potential. Biochemistry 49:1–10.

Morgan M, Iaconcig A, Muro AF. 2010. CPEB2, CPEB3 and CPEB4 are coordinately regulated by miRNAs recognizing conserved binding sites in paralog positions of their 3′-UTRs. Nucleic Acids Res 38:7698–7710.

Patil A, Kinoshita K, Nakamura H. 2010. Hub promiscuity in protein–protein interaction networks. Int J Mol Sci 11:1930–1943.

Prilusky J, Felder CE, Zeev-Ben-Mordehai T, Rydberg EH, Man O, Beckmann JS, Silman I, Sussman JL. 2005. FoldIndex: A simple tool to predict whether a given protein sequence is intrinsically unfolded. Bioinformatics 21:3435–3438.

Radivojac P, Vacic V, Haynes C, Cocklin RR, Mohan A, Heyen JW, Goebl MG, Iakoucheva LM. 2010. Identification, analysis, and prediction of protein ubiquitination sites. Proteins 78:365–380.

Rechsteiner M, Rogers SW. 1996. PEST sequences and regulation by proteolysis. Trends Biochem Sci 21:267–271.

Ren J, Gao X, Jin C, Zhu M, Wang X, Shaw A, Wen L, Yao X, Xue Y. 2009. Systematic study of protein sumoylation: Development of a site-specific predictor of SUMOsp 2.0. Proteomics 9:3409–3412.

Scanlan MJ, Simpson AJ, Old LJ. 2004. The cancer/testis genes: Review, standardization, and commentary. Cancer Immun 4:1.1–15.

Spassov DS, Jurecic R. 2002. Cloning and comparative sequence analysis of PUM1 and PUM2 genes, human members of the Pumilio family of RNA-binding proteins. Gene 299:195–204.

Stevenson BJ, Iseli C, Panji S, Zahn-Zabal M, Hide W, Old LJ, Simpson AJ, Jongeneel CV. 2007. Rapid evolution of cancer/testis genes on the X chromosome. BMC Genomics 8:129.1–11.

Suyama T, Shiraishi T, Zeng Y, Yu W, Parekh N, Vessella RL, Luo J, Getzenberg RH, Kulkarni P. 2010. Expression of cancer/testis antigens in prostate cancer is associated with disease progression. Prostate 70:1778–1787.

Tompa P, Csermely P. 2004. The role of structural disorder in the function of RNA and protein chaperones. FASEB J 18:1169–1175.

Uversky VN. 2002. Natively unfolded proteins: A point where biology waits for physics. Protein Sci 11:739–756.

Uversky VN, Dunker AK. 2010. Understanding protein non-folding. Biochim Biophys Acta 1804:1231–1264.

Uversky VN, Oldfield CJ, Dunker AK. 2008. Intrinsically disordered proteins in human diseases: Introducing the D2 concept. Annu Rev Biophys 37:215–246.

van Dieck J, Teufel DP, Jaulent AM, Fernandez-Fernandez MR, Rutherford TJ, Wyslouch-Cieszynska A, Fersht AR. 2009. Posttranslational modifications affect the interaction of S100 proteins with tumor suppressor p53. J Mol Biol 394:922–930.

Vavouri T, Semple JI, Garcia-Verdugo R, Lehner B. 2009. Intrinsic protein disorder and interaction promiscuity are widely associated with dosage sensitivity. Cell 138:198–208.

Welchman RL, Gordon C, Mayer RJ. 2005. Ubiquitin and ubiquitin-like proteins as multifunctional signals. Nat Rev Mol Cell Biol 6:599–609.

Westbrook VA, Schoppee PD, Diekman AB, Klotz KL, Allietta M, Hogan KT, Slingluff CL, Patterson JW, Frierson HF, Irvin WP Jr., Flickinger CJ, Coppola MA, Herr JC. 2004. Genomic organization, incidence, and localization of the SPAN-x family of cancer-testis antigens in melanoma tumors and cell lines. Clin Cancer Res 10:101–112.

Wiklund L, Sokolowski M, Carlsson A, Rush M, Schwartz S. 2002. Inhibition of translation by UAUUUAU and UAUUUUUAU motifs of the AU-rich RNA instability element in the HPV-1 late 3′ untranslated region. J Biol Chem 277:40462–40471.

Wong YH, Lee TY, Liang HK, Huang CM, Wang TY, Yang YH, Chu CH, Huang HD, Ko MT, Hwang JK. 2007. KinasePhos 2.0: A web server for identifying protein kinase-specific phosphorylation sites based on sequences and coupling patterns. Nucleic Acids Res 35:W588–W594.

Yang ZR, Thomson R, McNeil P, Esnouf RM. 2005. RONN: The bio-basis function neural network technique applied to the detection of natively disordered regions in proteins. Bioinformatics 21:3369–3376.

Zeng Y, He Y, Yang F, Mooney SM, Getzenberg RH, Orban J, Kulkarni P. 2011. The cancer/testis antigen prostate-associated gene 4 (PAGE4) is a highly intrinsically disordered protein. J Biol Chem 286:13985–13994.

Zhao R, Tang B, Liu Y, Zhu N. 2011. NLS-dependent and insufficient nuclear localization of XAGE-1 splice variants. Oncol Rep 25:1083–1089.