JPP JOURNAL OF PROTEINS AND PROTEOMICS
9(3), 2018, pp. 169-184



Research Article

# STRUCTURAL INTRICACY OF DISORDERED REGIONS IN TRANSCRIPTION FACTORS IMPARTING COLON CANCER

**Animesh Mondal[a], Uttam Pal[b], Anupam Roy[a] and Nakul C Maiti[a*]**

[a] *Structural Biology and Bioinformatics Division, Indian Institute of Chemical Biology, Council of Scientific and Industrial Research, 4, Raja S.C. Mullick Road, Kolkata 700032, India*
[b] *Chemical Sciences Division, Saha Institute of Nuclear Physics, 1/AF Bidhannagar, Kolkata, India*

***Abstract:*** Transcription factors (TFs) linked to cancer contains a significant amount of disordered segments and the dynamics as well as coupling of it with partner molecules are keys in the processes of recognization and many other cellular activities. However, the sequence distribution, composition and the structural adaptability of TFs in the presence of other interacting small and large macromolecules such as nucleic acids, lipids and other proteins are not well understood. In a recent article (*Specific DNA Sequences Allosterically Enhance Protein–Protein Interaction in a Transcription Factor, Phys. Chem. Chem. Phys.*) we showed that a small fluctuation in energetic and related conformational adaptation is a key in the recognition of multiple DNA sequences by TFs. In the current investigation we determined and showed the amino acid residue distribution, intrinsic characteristics, conformational adaptability of the disorder regions and their similarity among the TFs which are linked to colon cancer. About 38% of the residues in TFs in average found to belong in the disordered region (DoR) and the abundance of these DoRs follows a Poisson distribution pattern with an expectancy value of 12. Interestingly, the size (length) distribution of the individual DoR also follows a similar statistical pattern. The computational analysis further establishes the tertiary structure of many of the TFs by homology modeling and, it was observed that the TFs with disordered content less than 70% can attain different kind of tertiary folds. However, a significant segments preferred to remain as disordered (coiled coil) and found to localize preferentially on the surface of the proteins. This surface localization suggested interactive and functional links of the disordered regions with proteins and genomic materials involves in cell function and other biological activities. The presence and their localization may thus aid in the understanding of eukaryotic gene transcription machinery as well as developing new drugs targeting the disorder prone regions in the TFs to cure cancers and related human diseases.

***Keywords:*** TFs; Secondary structure; DoR; Functional link; PSIPRED

***Note:*** Colored figures and Supplementary information are available on the journal website

## Introduction

Recent investigations found that human proteome is enriched with intrinsically disorder regions (IDRs) in their sequence which in turn reflect their indispensable role in cellular signalling and various other metabolic pathways. The number of disorder proteins known to be involved in cell signalling and regulation is growing rapidly. Various well-characterized examples of individual disorder proteins involved in transcriptional regulation have been illustrated in literature (Dyson and Wright, 2002; Iakoucheva *et al*., 2002). However, the functional role of IDRs in crucial areas such as transcriptional regulation, translation and other

Corresponding Author: **Nakul C Maiti**
*E-mail: ncmaiti@iicb.res.in*

**Received**: August 31, 2018
**Accepted**: September 29, 2018
**Published**: September 30, 2018

cellular signal transduction has only been organized as a consequence of the use of new paradigms in biological methodology (DeForte and Uversky, 2017; Du and Uversky, 2017; Dyson and Wright, 2005). Intrinsically disorder region of the protein often involves in this recognization event. Several DNA binding proteins, including transcription factors (TFs) bind to DNA sequences however, with different affinity and the conformational adaptability (Banerjee and Chakraborty, 2017; Guo *et al*., 2017; Mazumder *et al*., 2017; Naiya *et al*., 2016; Uversky *et al*., 2014) of the protein-binding domain plays a key role in this protein-DNA interaction

Intrinsically disorder proteins (IDPs) or regions (IDRs) are characterized by the unique combination of high specificity and low affinity in their interaction with functional partners, which are very crucial for transient protein-protein, and protein-nucleic acid interactions, such as those that frequently occur during signal transduction, recognition and other cellular events. The propensity of disorder proteins or regions to form large interaction surfaces allowing them to wrap up or surround their binding partners (Liu *et al*., 2006). Disorder regions are essential for the function of transcriptional activators and numerous others signalling and regulatory proteins (Dyson and Wright, 2005). These IDRs are the site of many chromosomal translocations that are associated with the disease; for example translocation, that fuse region or CBP (CREB binding protein) or p300 to segments of MOZ (monocytic zinc finger leukemia protein) or MLL are associated with human leukemias (Goodman and Smolik, 2000; Yang, 2004). This phenomenon probably reflects the structural organization of proteins. Computational studies and experimental investigation further verified that binding regions in IDPs/IDRs are exposed and often considered as a primary contact site for the interaction and binding (Pal *et al*., 2016). In the present investigation, we aimed to procure the composition and conformational adaptability of the disorder regions in the transcription factors associated with colon cancer and derived some statistical correlation. Illustrating disorder regions and associated statistical knowledge is crucial to address functional and binding roles of proteins.

We have selected the transcription factors that are responsible for colon cancer from the extensive literature survey and 35 transcription factors were taken for our analysis purposes. Colon cancer is the third most common type of cancer in people. For deriving a new approach towards drug target analysis of DoR with in the TFs responsible for this cancer would be a novel stratagem.Various statistical models were enjoined to divulge the distribution of disorder region, hydrophobicity, and the structural preferences. Abundance (number) of disordered regions in the domain followed a Poisson distribution pattern with expectancy value of 12. Interestingly, the size (length) distribution of the individual DoR also followed a Poisson distribution. Conformational preferences of the disordered segments and the whole proteins were made by ab-initio method. It helped to trace the location of the disordered regions on the modeled structure; whether it is located on the inner surface or outer surface of the proteins. Interestingly, it was found that most of the disordered regions located on the outer surface of the proteins.

## Materials and Methods

### *Dataset Formation*

Transcription Factors (TFs) that are responsible for colon cancer were obtained from an extensive literature search and 35 TFs were chosen for our analysis purpose (Ballian *et al*., 2008; Bruun *et al*., 2014; Brzozowa *et al*., 2015; Cajuso *et al*., 2014; Cathomas, 2014; Caunt *et al*., 2015; Dawson *et al*., 2014; Francipane and Lagasse, 2013; Gerlach *et al*., 2012; Hackl *et al*., 2010; Hibi *et al*., 2009; Hu *et al*., 2010; Irby and Yeatman, 2000; Kobayashi *et al*., 1999; Li *et al*., 2014; Morishita *et al*., 2010; Naccarati *et al*., 2012; Network, 2012; Petrova *et al*., 2008; Slattery *et al*., 2010; Smith *et al*., 1993; Spano *et al*., 2005; Yang *et al*., 2018; Yao *et al*., 2013; Zhu *et al*., 2014). Sequences of these TFs were retrieved from UniProt. Sequences obtained from Uniprot in FASTA format were converted to strings of single letter amino acid codes for the further analysis purpose (Table 2).

### *Ab initio Modelling*

Structural model of TFs were rendered using I-TASSER (Iterative Threading ASSEmbly Refinement) services of Zhang Lab. It first identifies structural templates from the PDB by multiple threading approach LOMETS with full-length atomic model constructed by iterative template fragment assembly simulations.Further function insights of the target were then retrieved by threading 3D model through protein function database BioLiP (Roy *et al*., 2010).

### *Calculation of Disorder Regions*

Disorder residues and regions were identified using IUPred web server. It recognizes disorder region from the amino acid sequences of protein based on the estimated paired energy content. The assumption behind is that globular protein are composed of amino acids which have the tendency to form a large number of interactions. On the other hand, IDPs do not adopt any suitable structure because of their amino acid composition which creates hindrance in case of their interaction. This web server takes single amino acid sequences as an input and calculates the pairwise energy profile along the sequence. The energy value is then transformed into a probabilistic score ranging from 0 (complete order) to 1 (complete disorder) (Dosztányi *et al*., 2005). Residues above 0.5 regarded as disorder and the regions are counted by finding a stretch where is no ordered residue between the two consecutive disordered residues.

### *Analysis of Sequence*

The composition of amino acids, length, charged residues, total charge and molecular weight were calculated from the sequence of proteins. Gravy value of proteins and disorder regions were calculated as a result of hydropathy value of amino acids. Isoelectric points were calculated from ExPASy bioinformatics portal using ProtParam tool (Gasteiger *et al*., 2005). The secondary structural propensity of each protein for their amino acids were predicted using PSIPRED algorithm (Jones, 1999). Preference of a particular conformation of protein was measured by taking the ratio of total number of resiudes preferring particular conformation. Similarly, the conformation propensity of disorder residues and regions were derived from the mother protein using position and length parameter.

### *Functional Link*

Functional protein-protein interaction was determined by deriving the names of the associated proteins with our TFs.Proteins that are linked with TFs were fetched from String DataBase (Szklarczyk *et al*., 2015). Finally, all the proteins were compiled in Wolfram Mathematica 10. Interactions between them were determined by the number of their nodes i.e more nodes indicate more interactions.

### *Statistical Analysis*

All the statistical analysis was performed in Wolfram-Mathematica 10. Significance of the mean differences was established by performing Student's t-Test and the null hypothesis was rejected at the 5% level of significance.Poisson distribution was fitted to the Disorder region frequency (DoR) and length if the data.The Probability Mass Function is calculated by -

$$f(x;\mu) = \frac{e^{-\mu}\,\mu^{x}}{x!} \tag{1}$$

Where $e$ is Euler's number and μ is the expected value of the random variable $x$.

The Cumulative Function is calculated by –

$$g(x;\mu) = e^{-\mu}\,\Sigma_{i=0}^{[x]}\,\frac{\mu^{i}}{i!} \tag{2}$$

Where $[x]$ is the floor function.

Generalised Poisson distribution function is calculated by the equation –

$$f(x;\mu) = \frac{e^{-x\lambda-\mu}(x\lambda+\mu)^{-1+x}}{x!} \tag{3}$$

Where λ is a real number between 0 and 1.

Normal distribution with mean (μ) and standard deviation (σ) was fitted to the Normally distributed data.The probability distribution function is formulated through –

$$P(x) = \frac{1}{\sqrt{2}\pi\sigma}\,e^{\frac{(x-\mu)^2}{2\sigma^2}} \tag{4}$$

The Cumulative Distribution Function is calculated by –

$$D(x) = \frac{1}{2}\left[1 + erf\left[\frac{x-\mu}{\sqrt{2\sigma}}\right]\right] \tag{5}$$

Where *erf* is error function.

**Table 1**
**Poisson distribution parameters (λ) for DoR frequency and length of DoR**

| *Variable* | *λ* |
|---|---|
| DoR Frequency | 12.22 |
| DoR length | 22.26 |

**Table 2**
**Details of TFs and their disorder percentage**

| Sl No. | Protein Name | Uniprot ID | Localization | Function | Length | Disorder Seq. | Disorder % | PI | Hydro-pathy |
|---|---|---|---|---|---|---|---|---|---|
| 1 | Activin receptor type-1B isoform1 | UniProtKB - P36896 (ACV 1B_HUMAN) | Membrane | Forming an activin receptor complex with activin receptor type 2 and then interact with Smad. Helps in neuronal differentiation extracellular matrix production | 505 | 0 | 0 | 6.6 | -0.094 |
| 2 | Activin receptor type-2A isoform1 | UniProtKB - P27037 (AVR 2A_HUMAN) | Membrane | On ligand binding form a receptor complex and activate SMAD | 513 | 7 | 1.36 | 5.61 | -0.202 |
| 3 | AMER1 | UniProtKB - Q5JTC6 (AMER 1_HUMAN) | Cytoplasm/ Nucleus | Regulator of canonical Wnt signalling pathway | 1135 | 846 | 74.53 | 4.77 | -0.867 |
| 4 | Adenomatous polyposis coli protein | UniProtKB - P25054 (APC_ HUMAN) | Cytoplasm/ cell membrane | Tumor suppressor | 2843 | 2173 | 76.43 | 7.92 | -0.89 |
| 5 | ARID1A | UniProtKB - O14497 (ARI1A_ HUMAN) | Nucleus | Involves in chromatin remodelling | 2285 | 1755 | 76.80 | 6.24 | -0.778 |
| 6 | ATF3 | UniProtKB - P18847 (ATF3_ HUMAN) | Nucleus | Binds to cAMP response element | 181 | 57 | 31.49 | 8.8 | -0.671 |
| 7 | Serine/threonine-protein kinase B-raf | UniProtKB - P15056 (BRAF_ HUMAN) | Cell membrane/ cytoplasm/ Nucleus | Contribute to MAP Kinase signalling pathway | 766 | 219 | 28.59 | 7.29 | -0.36 |
| 8 | CDH3 | UniProtKB - P22223 (CADH3_ HUMAN) | Cell adherent junction, Cytosol | It Regulates beta catenin | 829 | 276 | 33.29 | 4.6 | -0.404 |
| 9 | Caspase-1 | UniProtKB - P29466 (CASP1_ HUMAN) | Cytoplasm | Promote apoptosis | 404 | 40 | 9.90 | 5.63 | -0.333 |
| 10 | CDX2 | UniProtKB - Q99626 (CDX2_ HUMAN) | Nucleus | Involved in transcriptional regulation of genes in intestinal epithelium | 313 | 205 | 65.49 | 9.65 | -0.674 |
| 11 | CTHRC1 isoform 1 | UniProtKB - Q96CG8 (CTHR1_ HUMAN) | Extra cellular matrix | Over expression causes invasion of CRC cells | 243 | 40 | 16.46 | 8.31 | -0.282 |

*contd. table 2*

| Sl No. | Protein Name | Uniprot ID | Localization | Function | Length | Disorder Seq. | Disorder % | PI | Hydro-pathy |
|---|---|---|---|---|---|---|---|---|---|
| 12 | Receptor tyrosine-protein kinase erbB-2 | UniProtKB - P04626 (ERBB2_ HUMAN) | Cytoplasm/ Nucleus | Plays role in RTK pathway | 1255 | 244 | 19.44 | 5.58 | -0.247 |
| 13 | Receptor tyrosine-protein kinase erbB-3 | UniProtKB - P21860 (ERBB3_ HUMAN) | Plasma membrane | Transmembrane signalling molecule may interact with ERBB2 | 1342 | 353 | 26.30 | 6.11 | -0.387 |
| 14 | MEK 1 | UniProtKB - Q02750 (MP2K1_ HUMAN) | Cytoplasm/ Nucleus | Regulation of Map kinase signal Transduction pathway | 393 | 63 | 16.03 | 6.18 | -0.305 |
| 15 | Serine/threonine-protein kinase mTOR | UniProtKB - P42345 (MTOR_ HUMAN) | Cytoplam /Nucleus/ER/ Mitochondria | Activated mTORC control protein synthesis by phospho regulating key factor | 2549 | 554 | 21.73 | 6.73 | -0.193 |
| 16 | Myc proto-oncogene protein (MYC) | UniProtKB - P01106 (MYC_ HUMAN) | Nucleus | TF activates growth related genes | 439 | 208 | 47.38 | 5.33 | -0.772 |
| 17 | NFE2L3 | UniProtKB - Q9Y4A8 (NF2L3_ HUMAN) | Nucleus | Activates erythroid specific globin gene expression | 694 | 333 | 47.98 | 5.21 | -0.627 |
| 18 | NFATc2 | UniProtKB - Q13469 (NFAC 2_HUMAN) | Cytoplasm/ Nucleus | Involves in expression of cytokine gene in T Cells and promotes invasive migration through GPC6 expression and Wnt signalling | 925 | 513 | 55.45 | 6.87 | -0.544 |
| 19 | P53 | UniProtKB - P04637 (P53_ HUMAN) | Cytolplasm/ Nucleus | Tumor suppressor | 393 | 210 | 53.43 | 6.33 | -0.756 |
| 20 | PDX1 | UniProtKB - P52945 (PDX1_ HUMAN) | Nucleus/ Cytoplasm | Glucose dependent transcription regulation | 283 | 96 | 33.92 | 7.1 | -0.671 |
| 21 | PIK3CA | UniProtKB - P42336 (PK3CA_ HUMAN) | Cytosol | Involved in AKT activation and cell proliferation,growth | 1068 | 17 | 1.59 | 6.88 | -0.306 |
| 22 | PLAC8 (PLACENTA SPECIFIC 8) | UniProtKB - Q96EJ4 (Q96EJ4_ HUMAN) | Cytosol (Intestine) | It regulates DUSP6 which in turn controls p-ERK2 | 115 | 0 | 0 | 7.34 | 0.125 |

*contd. table 2*

| Sl No. | Protein Name | Uniprot ID | Localization | Function | Length | Disorder Seq. | Disorder % | PI | Hydro-pathy |
|---|---|---|---|---|---|---|---|---|---|
| 23 | PROX1 | UniProtKB - Q92786 (PROX1_ HUMAN) | Nucleus | Plays a role in embryonic development | 737 | 432 | 58.61 | 6.74 | -0.757 |
| 24 | RAC-alpha serine/threonine-protein kinase | UniProtKB - P31749 (AKT1_ HUMAN) | Cytoplasm/ Nucleus/Cell membrane | It is a kinase regulates cell growth, metabolism,apoptosis and MAP3K5 | 480 | 90 | 18.75 | 5.75 | -0.575 |
| 25 | Mothers against decapentaplegic homolog 2 isoform-Long | UniProtKB - Q15796 (SMAD2_ HUMAN) | Cytoplasm/ Nucleus | Intracellular signal transducer & transcriptional modulator activated by TGF-Beta | 467 | 88 | 18.84 | 6.13 | -0.444 |
| 26 | Mothers against decapentaplegic homolog 3 isoform1 | UniProtKB - P84022 (SMAD3_ HUMAN) | Cyoplasm/ Nucleus | Component of heteromeric complex required for Tgf-Beta signalling | 425 | 83 | 19.52 | 6.73 | -0.447 |
| 27 | Mothers against decapentaplegic homolog 7 | UniProtKB - O15105 (SMAD7_ HUMAN) | Cyoplasm/ Nucleus | Antagonist of signalling by TGF-Beta | 426 | 90 | 21.12 | 8.63 | -0.398 |
| 28 | SNAI1 | UniProtKB - O95863 (SNAI1_ HUMAN) | Nucleus/ Cytoplasm | Induction of EMT | 264 | 74 | 28.03 | 8.97 | -0.516 |
| 29 | SOX9 | UniProtKB - P48436 (SOX9_ HUMAN) | Nucleus | Facilitate beta catenin degradation | 509 | 449 | 88.21 | 6.31 | -1.007 |
| 30 | Protooncogene tyrosine protein kinase SRC (SRC) | UniProtKB - P12931 (SRC_ HUMAN) | Cell membrane/ Nucleus/ Cytoplasm | Primary kinase plays role in cytoskeletal reorganization | 536 | 82 | 15.29 | 7.1 | -0.473 |
| 31 | Stimulator of interferon genes protein (STING) | UniProtKB - Q86WV6 (STING_ HUMAN) | Cytoplasmic vesicle membrane/ Endoplasmic reticulum/ Goldi/ Mitochondria | Innate immunity | 379 | 45 | 11.87 | 6.6 | -0.054 |

*contd. table 2*

| Sl No. | Protein Name | Uniprot ID | Localization | Function | Length | Disorder Seq. | Disorder % | PI | Hydro-pathy |
|---|---|---|---|---|---|---|---|---|---|
| 32 | TGF-beta receptor type-2 isoform1 | UniProtKB - P37173 (TGFR2_ HUMAN) | Cytosol/ Cell membrane | Receptor of Tgf-beta | 567 | 17 | 2.99 | 5.6 | -0.354 |
| 33 | TGF-beta receptor type-1 | UniProtKB - P36897 (TGFR1_ HUMAN) | Cell membrane/ Tight junction | Receptor of Tgf-beta | 503 | 2 | 0.39 | 7.51 | -0.055 |
| 34 | Transforming growth factor beta-1 | UniProtKB - P01137 (TGFB1_ HUMAN) | Extra cellular matrix/ Space | Controls cell proliferation, differentiation, invasion and also connected with SMAD | 390 | 21 | 5.38 | 8.83 | -0.311 |
| 35 | Protein Wnt-5a | UniProtKB - P41221 (WNT5A_ HUMAN) | Extra cellular matrix | Ligands for members of the frizzeled family | 380 | 0 | 0 | 8.83 | -0.29 |

**Result and Discussion**

Transcription factors (TFs) are involved in different cell regulatory pathways. To understand their regulatory diversification in the purview of the structural aspect we analyzed the TFs sequences of the dataset. We used IUPred algorithm to determine disorder residues and regions. The analysis showed variation in TFs degree of disorderness and functions as given in (Table 1). The disorder regions (DoRs) were selected from the protein sequence based on the estimated paired energy content. Most of the proteins contained disorder regions. (Table 2) gives the details of the disorder region and the parent proteins. IUPred detected total 428 disorder regions with varying degree of disorderness and they were distributed with in 35 number of TFs.

The occurrence (frequency distribution) pattern of the DoRs present in the TFs showed a poisson distribution pattern (Fig. 1) and indicated that the occurrence of the regions was a stochastic in nature and it satisfied the Markov Property (Durbin *et al*., 1998). The Poisson distribution always provides the expectation value (λ) and, the expected occurrence (λ, statistically obtained value) of disordered regions was found to be 12. The sizes (length) of the diorder regions also followed a poisson distribution pattern (Fig. 1) and the expectation value for the length was close to 22. Fig. 2 shows the regression analysis to derive a correlation between protein length and the content of residues in the DoRs. It was found that the percentage of the disordered residues and the length of the protein was poorly correlated to each other with a very low $R^2$ value of 0.15. This conferred that the protein disordereness was independent of protein length.

Analysis was also made to detail the content of different amino acid residues in the DoRs and in total TFs. The content of different amino acid residues in the composition of the DoRs and TFs are shown in (Fig. 3). Panel A in the figure compares the sequence composition of DoRs with the Total TFs. Although the distribution pattern of amino acids for both the two was similar, an apparent difference in the sequence population of amino acid in TFs was observed Panel B in (Fig. 3). Residues S, Q, G and P were at least 1% higher in number than in the total protein sequence. It was expected that the disordered segments will be rich with G, P residues. However, also we observed higher amount of E,T, N and A in the disordered regions. Hydrophobic amino acids such as G and A are also abundant in the disorder prone region of TFs. Glycine is a unique amino acid because it contains hydrogen as its side chain other than carbon containing group in case of other amino acids. This confers much more flexibility when TFs interact with other partner. Moreover, it can reside in part of protein structures when it forms turn where all other amino acids are disallowed. Alanine is a non polar amino acid and its side chain is non reactive but it can play role in substrate recognition. These residues are with strong helical propensity (Koehl and Levitt, 1999) and may helps in attaining ordered structure upon binding with a suitable partner. It was also observed that V, I, L, F residues were specifically decreased in the disorder region in comparison to the total protein sequences. Interestingly the abundance of C residues was less; being TFs the proteins needs more pliability and evolutionary such selection was required. From our previous report (Das *et al*., 2014), it was well established that low complexity region which is also responsible for disorderness, promoted by P, G, E, S, Q and A residues. While analyzing the distribution pattern of the previously mentioned amino acids we have found them in high number within the disordered regions of TFs (Fig. 3B). Additionally, the residues that are responsible for promoting compact structure like C and W increased and P decreased, but the rest of the residues remain almost unchanged in the three classes. In case of MDP and LDP, there was a distinct difference in the number of G residues. LDPs are highly flexible in nature and for satisfying that feature high number of G and a low number of M and V is highly justifiable. The whole TFs showed amino acid distribution pattern similar to the total human disorder proteins found in DisProt database (Fig. S1). Athough the abundance of K and V amino acids were higher in TFs compared to the DisProt proteins but P and S were lower in case of TFs.

From various investigation, it has been well documented that tight regulation of IDP is very necessary for response to specific stimuli (Babu, 2016). In our study we have predicted the possible link up among the TFs by deriving its interacting partner from String Database (Fig. 4). In the predicted link these TFs showed maximum interspecific interaction between the TFs. However,CASP1 and PROX1 preferred maximum intraspecific interaction. Functionally CASP1 is very important because it is the key player in apoptosis pathway and PROX1 is a specific target of the â-

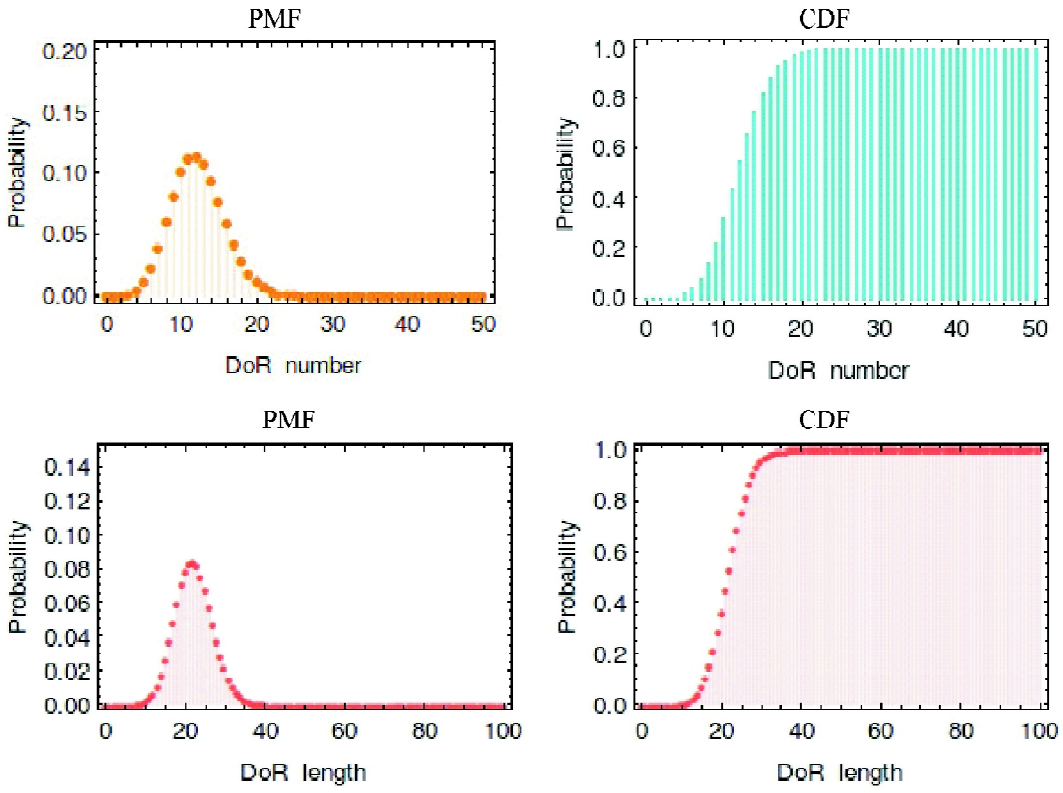



*Figure 1:* **Frequency and length distribution of the DoR. (A) Probability of the occurence of DoR in TFs.Probability mass function (PMF) of the fitted Poisson distribution is shown. (B) Cumulative distribution function of the DoR frequency. (C) PMF of individual DoR length in TFs.(D) CDF of the DoR length distribution**

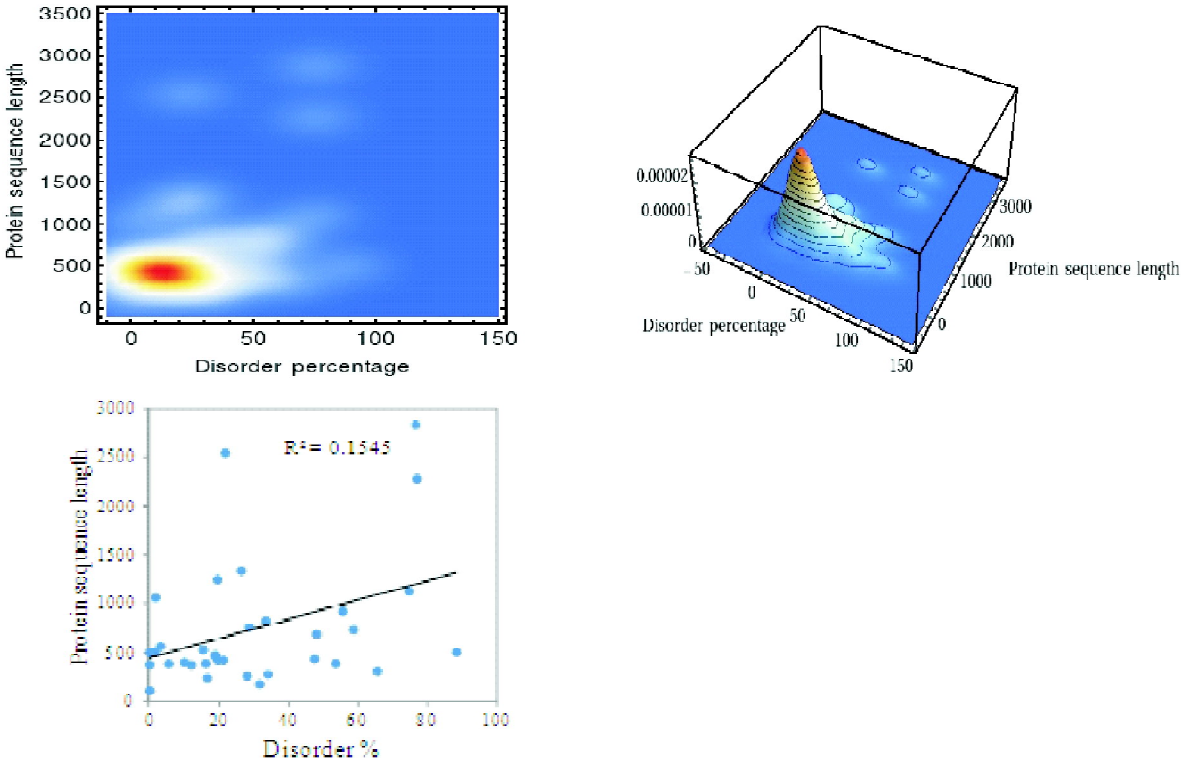



*Figure 2 :* **Correlation between the Protein length (TFs) and disorder % in TFs. (A) Distribution of the protein disorderness across its length. (B) Fitted linear model of the disorder % and length distribution**

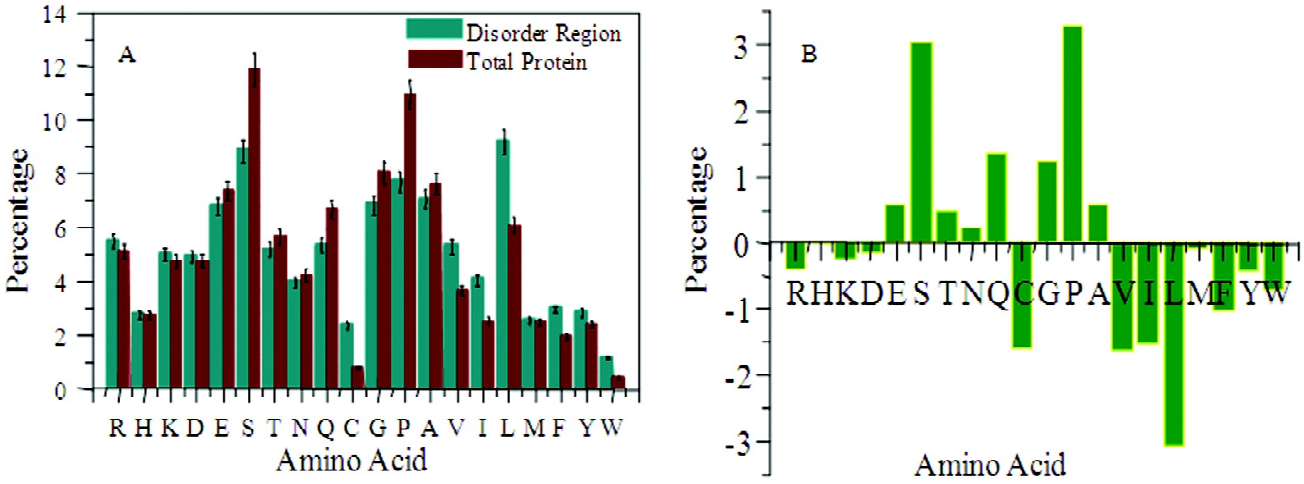



*Figure 3 :* **(A) Comparison of Amino Acid distribution between Whole TFs and DoR. Here the difference in distribution of amino acids were considered significant at P value <0.05**
**(B) Difference of amino acid between disorder region and total protein**

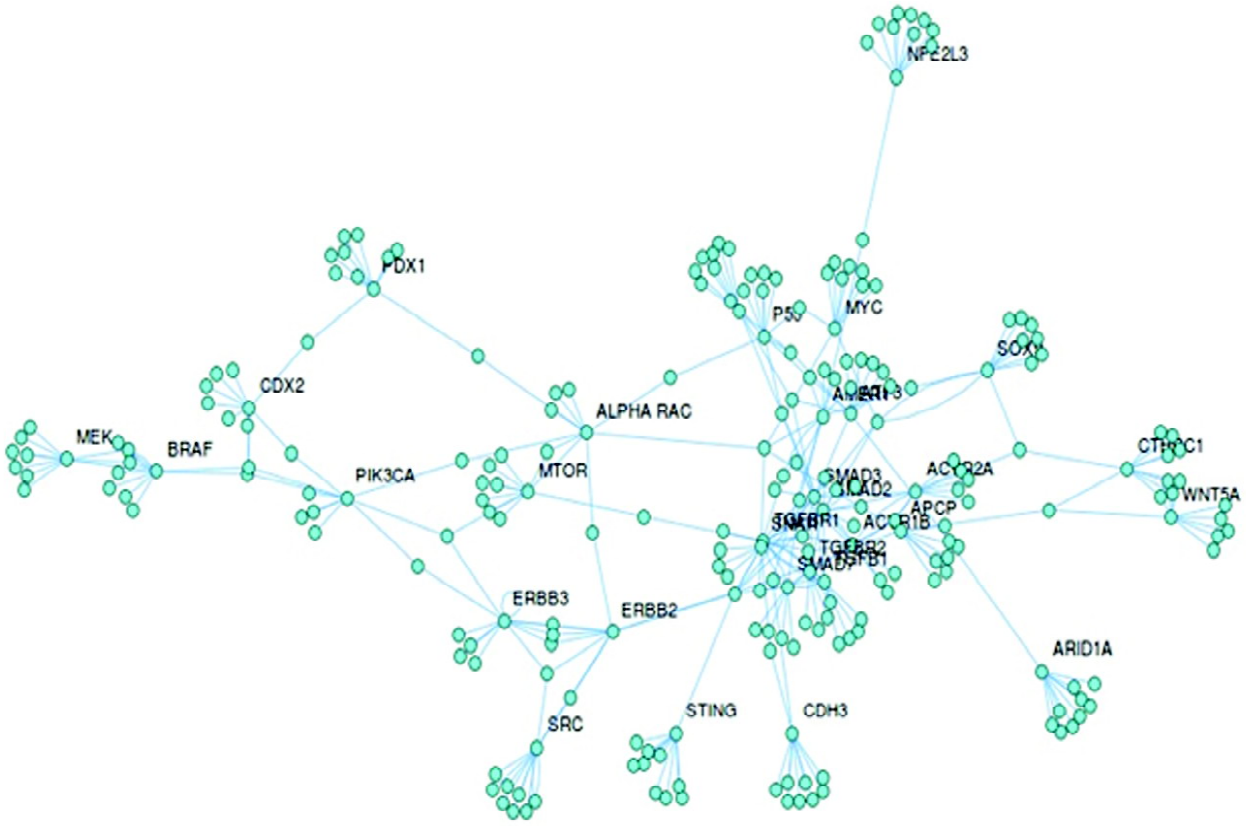



*Figure 4 :* **Predicted functional link among TFs.TFs acts via different other factors which are also inter linked with each other. Here probable link between these TFs are indicated**

Catenin/TCF pathway. From the prospect of disorderness CASP1 and PROX fall under PDP and MDP group respectively.

As interaction and functionality of the DoRs are of significant importance the hydrophobicity of the regions were judged by measuring the hydropathy indexes. Hydropathy index is the most preferable tool (Wu *et al*., 2006) that provide much intricate results about the hydrophobicity. By using hydropathy indexes of each amino acids the grand average hydropathy (GRAVY) values of DoR and the parent TFs were determined (Table 3). The calculated GRAVY indexes of majority of the TFs were predominantly negative and varied between 0 to -2. However, in case of the DoRs it varied between -4.5 to +4.5 (Fig. 5 and 6). Generally, if the hydropathy score lies below zero it confers the proteins most likely to be globular, where as hydropathy score above zero indicates membranous nature of the proteins.

Revealing disorder regions and its structural preference in TFs could unveil the functional peculiarity of TFs and the secondary structural preferences of the residues were calculated by IUPred algorithm. The amino acid residues, both in the TFs and DoRs within them prefer to adopt all the three major conformations such as helix, strand and coil which are depicted in the (Fig. 7). IUPred analysis is independent of secondary structure (Dosztányi *et al*., 2005) and therefore the predicted structure may not be largely biased. The analysis revealed less strand/beta sheet conformation in the DoRs (Fig. 8). Preferences for coiled conformation was found to be most abundant in the DoRs and some of the residues prefer to adopt helix conformation. Also, as expected, the structural propensity of DoR towards coil conformation was reasonably higher than the parental TFs. The overall structural content of TFs sequence was ~22% helix, 7% strand/sheet and 70% coil indicating the flexible nature of the TFs. However, in the DoR, the conformational preference was ~11% helix, 2% strand and 86% coil conformation. This indicated that a good number of the DoR residues potentially can be converted into alpha helical structural domain under suitable solution condition or in the presence of other molecule.

We further determined the molecular structure (2D and 3D) of the TFs by homology modeling (Fig. 9 and Fig. S2) and investigated the status of the DoRs in the modeled 3D representation. As expected, the TFs with lesser (less than 70%) disordered sequences yield 3D structures rich in alpha helical folds. The secondary structural propensity of the DoRs quite matches with the Psipred predicted secondary structural propensity of the residues in the region. For instance, ATF3

**Table 3**
**Fitted Normal distribution parameters for GRAVY**

| *Group* | $\mu$ | $\sigma$ |
|---|---|---|
| Protein (TFs) | -0.45 | 0.26 |
| DoR | -0.615 | 1.85 |

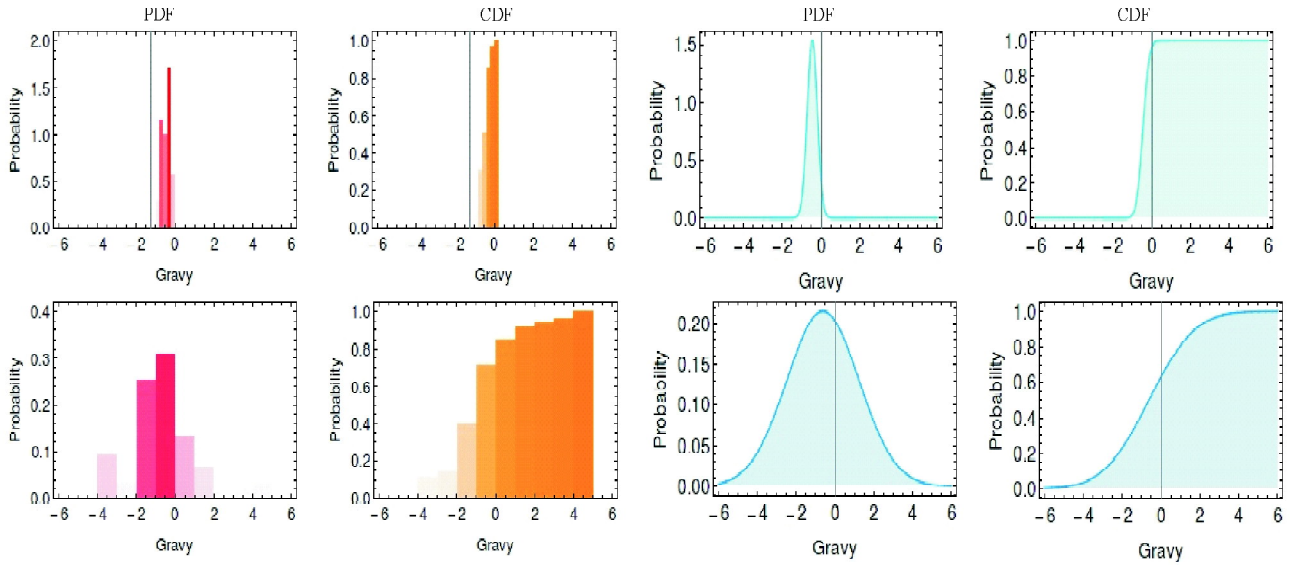



*Figure 5 :* **GRAVY distribution of the Whole TFs and DoR. Row1: GRAVY of the whole TFs.Row2: GRAVY of the DoR. Column 1&2: Histogram PDF & CDF. Column3&4: Fitted Normal distribution of the PDF & CDF**

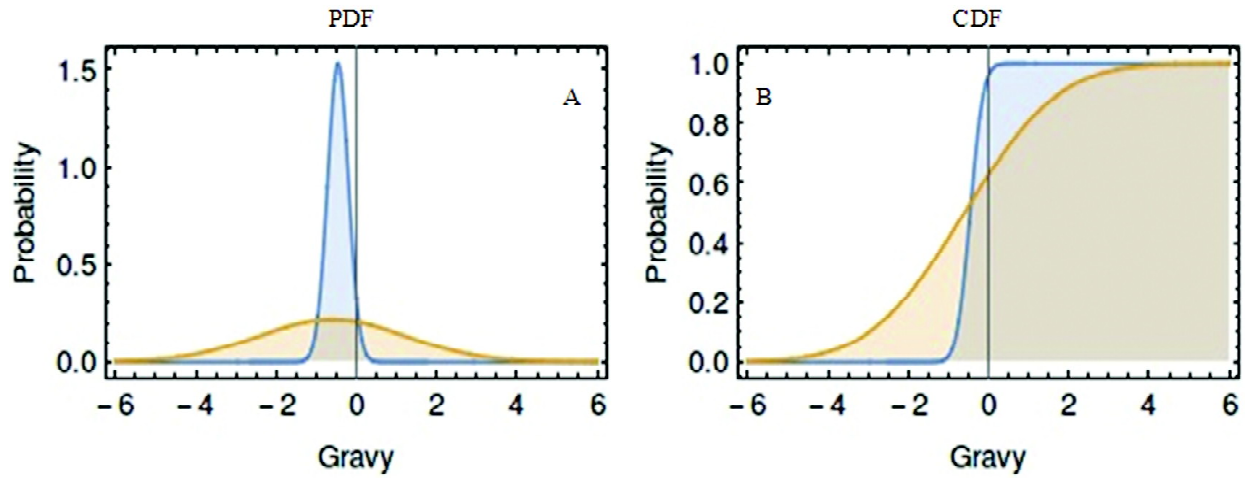



*Figure 6:* **Comparison of GRAVY between Whole TFs and DoR. (A ) PDF,(B) CDF. Color Key: Light Blue – Whole TFs, Grey-DoR**

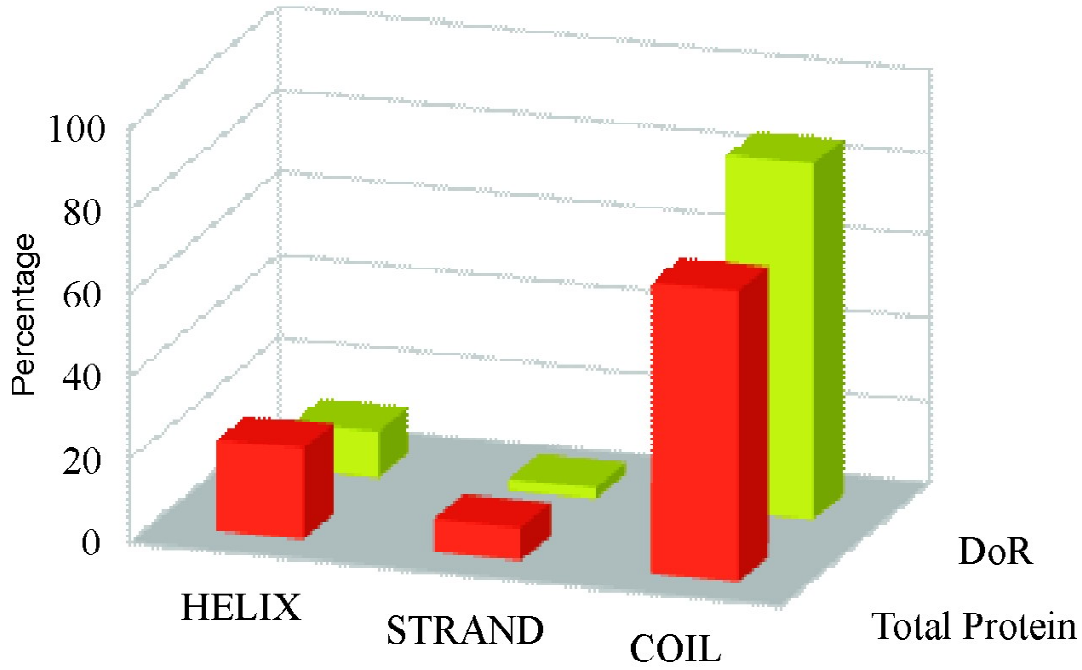



*Figure 7:* **Distribution of Secondary structure in TFs and DoR.Here Coil conformation(86%) in highly preferred among the disorder regions, followed by helix and strand conformation**

(30% of residues belong to disordered regions) the longest disordered region, residues from 76 to 102 showed structural propensity mainly to helix and coiled conformation that largely contribute to the disorderness. In the homology modeling structure also it showed similar structural propensity. Another important observation was that the disordered segments obtained by homology modeling remained mostly on the surface (Fig. S3) or stayed as flanks and extended to the solvent. However, large disordered regions, particularly in the middle of the protein sequence disallow the protein to adopt any compact and big folded structure. For instance, the disordered segment comprising of residues 70-89 in NFAC2 showed an extended structure in its 3D modeled structure (Fig. S2). ERBB2, also in its tertiary model reveals elongated conformation due to the presence of a long disorder segment. Amusingly, in case of CDX2 although it carries a long DoR of residues (70-90), its tertiary structure was not much elongated and contain significant amount of globular folded structure.

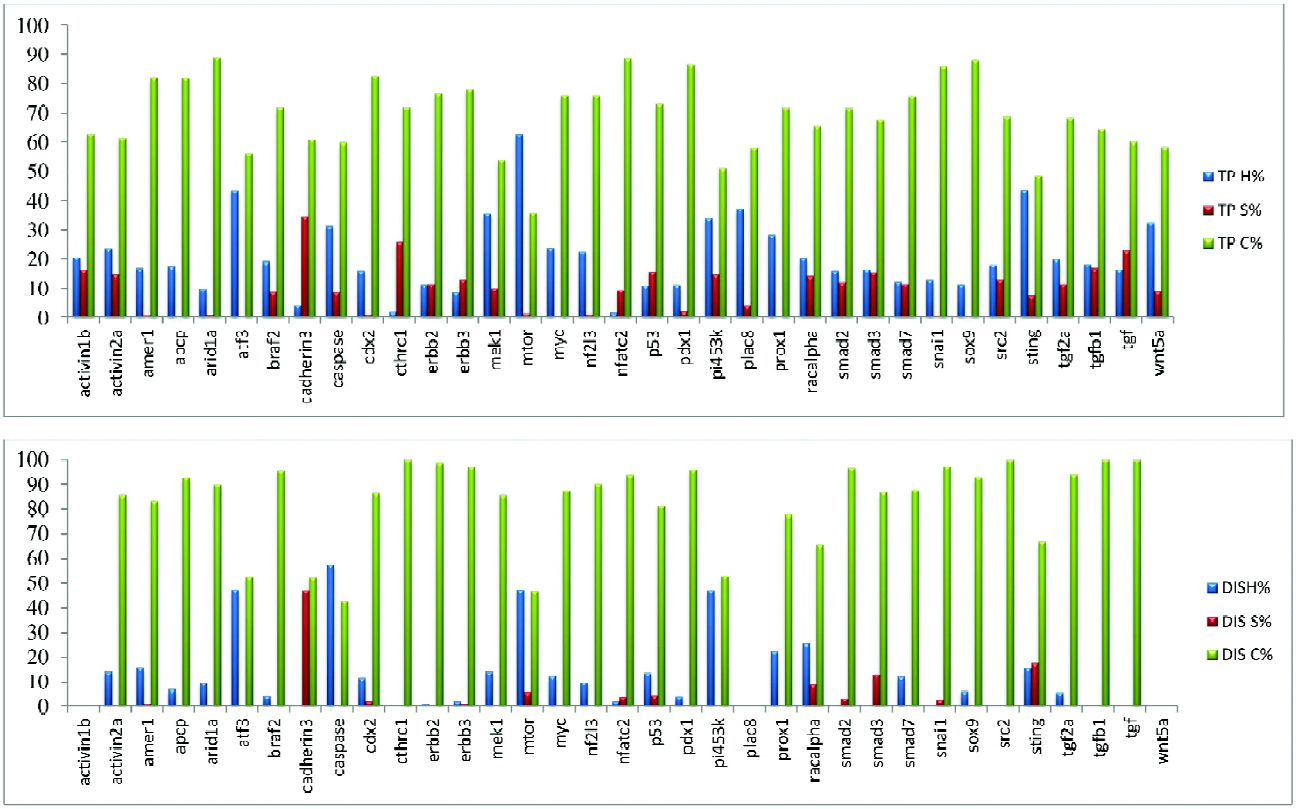



*Figure 8 :* **Comparison of Secondary Structutre conformation in TFs and with in the disorder region (DoR)**

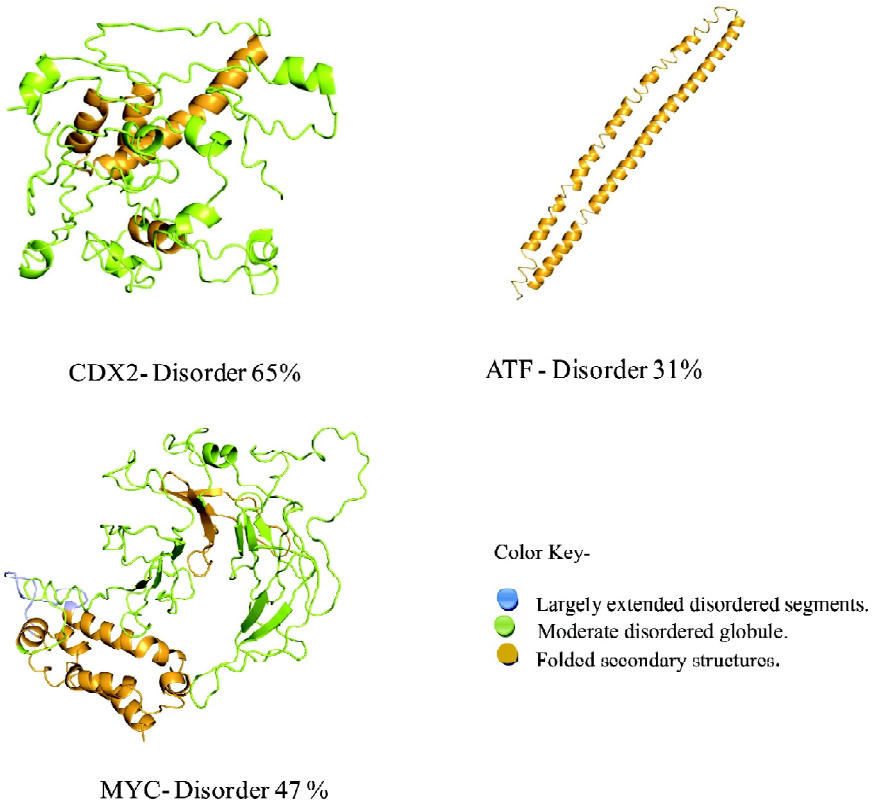



*Figure 9:* **Predicted model of TFs along with its disorder region (DoR) of CDX2 and NFAC2 . Predicted 3D conformation Of Transcription factors.**

## Conclusion

In the current manuscript we discussed the sequence aspects and the distribution pattern of disordered regions in TFs associated with colon cancer. It is reported that transcription factors in general contain intrinsically disordered regions in their sequence and contribute to their transactivation activity. It was reported that more than 80% of all transcription factors contain regions that are intrinsically disordered. Highly disorder proteins do not adopt any stable structure because of their amino acid composition which mostly disfavor hydrophobic collapse. In addition favorable solvation energy creates hindrance in formation of ordered and compact structure. The presence of disorderness in protein sequences reflects their greater need for cooperation, coordination and liaisons in diverse signalling pathway. Here, we investigated the disorder in transcription factors that are associated with colon cancer. Through a critical examination of disorder region by combining statistical and physiochemical information we determined the occurrence of disorder region in TFs and showed the distribution of disorder region length and its frequency in the parent TFs. This disorder region showed amphipathic nature. In the modeled 3D structure the disordered region found to prefer localization on the surface area of the protein. Such localization may significantly enhance useful in the initial stages of target selection for drug molecules. It is well known that the presence of well defined three-dimensional structure is the prerequisite of protein function (Tompa, 2003). Majority of the proteins used to adopt a defined three-dimensional structure to carry out their function. Therefore, the drug targets are often chosen in this structured regions. The development of a new advanced way to discover drug molecule would be interesting if this disorder region in TFs are taken into account.

### *Acknowledgement*

Animesh Mondal thanks, CSIR,Govt. of India for research fellowship. Anupam Roy, thanks UGC for the research fellowship. The authors thank Sandip Dolui for giving valuable comment on the manuscript.

***Conflict of Interest:*** The authors declare no competiting financial interests.

### *Abbreviations*

TFs: Transcription Factors; DoR: Disorder Region; IDPs: Intrinsically disorder proteins; IDRs: Intrinsically disorder regions.

### *Supporting information*

Figure - S1 Comparison of distribution of amino acid with in the TFs and total human protein found in DisProt Data base; Figure - S2 Prdicted 3D conformation Of Transcription factors and Figure - S3 Predicted model structure of TFs exhibiting DoRs.

### *References*

Babu, M.M. (2016). The contribution of intrinsically disordered regions to protein function, cellular complexity, and human disease. Biochem. Soc. Trans. *44*, 1185–1200.

Ballian, N., Liu, S.-H., and Brunicardi, F.C. (2008). Transcription factor PDX-1 in human colorectal adenocarcinoma: a potential tumor marker? World J. Gastroenterol. *14*, 5823–5826.

Banerjee, S., and Chakraborty, S. (2017). Protein intrinsic disorder negatively associates with gene age in different eukaryotic lineages. Mol. Biosyst. *13*, 2044–2055.

Bruun, J., Kolberg, M., Nesland, J.M., Svindland, A., Nesbakken, A., and Lothe, R.A. (2014). Prognostic Significance of â-Catenin, E-Cadherin, and SOX9 in Colorectal Cancer: Results from a Large Population-Representative Series. Front. Oncol. *4*, 118.

Brzozowa, M., Michalski, M., Wyrobiec, G., Piecuch, A., Dittfeld, A., Harabin-S³owiñska, M., Boroñ, D., and Wojnicz, R. (2015). The role of Snail1 transcription factor in colorectal cancer progression and metastasis. Contemp. Oncol. Poznan Pol. *19*, 265–270.

Cajuso, T., Hänninen, U.A., Kondelin, J., Gylfe, A.E., Tanskanen, T., Katainen, R., Pitkänen, E., Ristolainen, H., Kaasinen, E., Taipale, M., *et al*. (2014). Exome sequencing reveals frequent inactivating mutations in ARID1A, ARID1B, ARID2 and ARID4A in microsatellite unstable colorectal cancer. Int. J. Cancer *135*, 611–623.

Cathomas, G. (2014). PIK3CA in Colorectal Cancer. Front. Oncol. *4*,35.

Caunt, C.J., Sale, M.J., Smith, P.D., and Cook, S.J. (2015). MEK1 and MEK2 inhibitors and cancer therapy: the long and winding road. Nat. Rev. Cancer *15*, 577–592.

Das, S., Pal, U., Das, S., Bagga, K., Roy, A., Mrigwani, A., and Maiti, N.C. (2014). Sequence Complexity of Amyloidogenic Regions in Intrinsically Disordered Human Proteins. PLOS ONE *9*, e89781.

Dawson, H., Galván, J.A., Helbling, M., Muller, D.-E., Karamitopoulou, E., Koelzer, V.H., Economou, M., Hammer, C., Lugli, A., and Zlobec, I. (2014). Possible role of Cdx2 in the serrated pathway of colorectal cancer characterized by BRAF mutation, high-level CpG Island methylator phenotype and mismatch repair-deficiency. Int. J. Cancer *134*, 2342–2351.

DeForte, S., and Uversky, V.N. (2017). Not an exception to the rule: the functional significance of intrinsically disordered protein regions in enzymes. Mol. Biosyst. *13*, 463–469.

Dosztányi, Z., Csizmok, V., Tompa, P., and Simon, I. (2005). IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. Bioinforma. Oxf. Engl. *21*, 3433–3434.

Du, Z., and Uversky, V.N. (2017). Functional roles of intrinsic disorder in CRISPR-associated protein Cas9. Mol. Biosyst. *13*, 1770–1780.

Durbin, R., Eddy, S., Krogh, A., and Mitchison, G. (1998). Biological sequence analysis: probabilistic models of proteins and nucleic acids.

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. Nat. Rev. Mol. Cell Biol. *6*, 197–208.

Francipane, M.G., and Lagasse, E. (2013). mTOR pathway in colorectal cancer: an update. Oncotarget *5*, 49–66.

Gasteiger, E., Hoogland, C., Gattiker, A., Duvaud, S., Wilkins, M.R., Appel, R.D., and Bairoch, A. (2005). Protein Identification and Analysis Tools on the ExPASy Server. In The Proteomics Protocols Handbook, (Humana Press), pp. 571–607.

Gerlach, K., Daniel, C., Lehr, H.A., Nikolaev, A., Gerlach, T., Atreya, R., Rose-John, S., Neurath, M.F., and Weigmann, B. (2012). Transcription factor NFATc2 controls the emergence of colon cancer associated with IL-6-dependent colitis. Cancer Res. *72*, 4340–4350.

Goodman, R.H., and Smolik, S. (2000). CBP/p300 in cell growth, transformation, and development. Genes Dev. *14*, 1553–1577.

Guo, X., Han, J., Luo, R., and Chen, H.-F. (2017). Conformation dynamics of the intrinsically disordered protein c-Myb with the ff99IDPs force field. RSC Adv. *7*, 29713–29721.

Hackl, C., Lang, S.A., Moser, C., Mori, A., Fichtner-Feigl, S., Hellerbrand, C., Dietmeier, W., Schlitt, H.J., Geissler, E.K., and Stoeltzing, O. (2010). Activating transcription factor-3 (ATF3) functions as a tumor suppressor in colon cancer and is up-regulated upon heat-shock protein 90 (Hsp90) inhibition. BMC Cancer *10*, 668.

Hibi, K., Goto, T., Mizukami, H., Kitamura, Y.-H., Sakuraba, K., Sakata, M., Saito, M., Ishibashi, K., Kigawa, G., Nemoto, H., *et al.* (2009). Demethylation of the CDH3 gene is frequently detected in advanced colorectal cancer. Anticancer Res. *29*, 2215–2217.

Hu, B., Elinav, E., Huber, S., Booth, C.J., Strowig, T., Jin, C., Eisenbarth, S.C., and Flavell, R.A. (2010). Inflammation-induced tumorigenesis in the colon is regulated by caspase-1 and NLRC4. Proc. Natl. Acad. Sci. U. S. A. *107*, 21635–21640.

Iakoucheva, L.M., Brown, C.J., Lawson, J.D., Obradoviæ, Z., and Dunker, A.K. (2002). Intrinsic Disorder in Cell-signaling and Cancer-associated Proteins. J. Mol. Biol. *323*, 573–584.

Irby, R.B., and Yeatman, T.J. (2000). Role of Src expression and activation in human cancer. Oncogene *19*, 5636–5642.

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

Kobayashi, A., Ito, E., Toki, T., Kogame, K., Takahashi, S., Igarashi, K., Hayashi, N., and Yamamoto, M. (1999). Molecular cloning and functional characterization of a new Cap'n' collar family transcription factor Nrf3. J. Biol. Chem. *274*, 6443–6452.

Koehl, P., and Levitt, M. (1999). Structure-based conformational preferences of amino acids. Proc. Natl. Acad. Sci. *96*, 12524–12529.

Li, C., Ma, H., Wang, Y., Cao, Z., Graves-Deal, R., Powell, A.E., Starchenko, A., Ayers, G.D., Washington, M.K., Kamath, V., *et al.* (2014). Excess PLAC8 promotes an unconventional ERK2-dependent EMT in colon cancer. J. Clin. Invest. *124*, 2172–2187.

Liu, J., Perumal, N.B., Oldfield, C.J., Su, E.W., Uversky, V.N., and Dunker, A.K. (2006). Intrinsic disorder in transcription factors. Biochemistry (Mosc.) *45*, 6873–6888.

Mazumder, A., Batabyal, S., Mondal, M., Mondol, T., Choudhury, S., Ghosh, R., Chatterjee, T., Bhattacharyya, D., Pal, S.K., and Roy, S. (2017). Specific DNA sequences allosterically enhance protein–protein interaction in a transcription factor through modulation of protein dynamics: implications for specificity of gene regulation. Phys. Chem. Chem. Phys. *19*, 14781–14792.

Morishita, A., Gong, J., Nomura, T., Yoshida, H., Izuishi, K., Suzuki, Y., Kushida, Y., Haba, R., D'Armiento, J., and Masaki, T. (2010). The use of protein array to identify targetable receptor tyrosine kinases for treatment of human colon cancer. Int. J. Oncol. *37*, 829–835.

Naccarati, A., Polakova, V., Pardini, B., Vodickova, L., Hemminki, K., Kumar, R., and Vodicka, P. (2012). Mutations and polymorphisms in TP53 gene—an overview on the role in colorectal cancer. Mutagenesis *27*, 211–218.

Naiya, G., Raha, P., Kumar Mondal, M., Pal, U., Saha, R., Chaudhuri, S., Batabyal, S., Pal, S.K., Bhattacharyya, D., C. Maiti, N., *et al.* (2016). Conformational selection underpins recognition of multiple DNA sequences by proteins and consequent functional actions. Phys. Chem. Chem. Phys. *18*, 21618–21628.

Network, T.C.G.A. (2012). Comprehensive molecular characterization of human colon and rectal cancer. Nature *487*, 330–337.

Pal, U., Maity, M., Khot, N., Das, S., Das, S., Dolui, S., and Maiti, N.C. (2016). Statistical insight into the binding regions in disordered human proteome. J. Proteins Proteomics 7,47-60.

Petrova, T.V., Nykänen, A., Norrmén, C., Ivanov, K.I., Andersson, L.C., Haglund, C., Puolakkainen, P., Wempe, F., von Melchner, H., Gradwohl, G., *et al.* (2008). Transcription factor PROX1 induces colon cancer progression by promoting the transition from benign to highly dysplastic phenotype. Cancer Cell *13*, 407–419.

Roy, A., Kucukural, A., and Zhang, Y. (2010). I-TASSER: a unified platform for automated protein structure and function prediction. Nat. Protoc. *5*, 725–738.

Slattery, M.L., Herrick, J.S., Lundgreen, A., Fitzpatrick, F.A., Curtin, K., and Wolff, R.K. (2010). Genetic variation in a metabolic signaling pathway and colon and rectal cancer risk: mTOR, PTEN, STK11, RPKAA1, PRKAG2, TSC1, TSC2, PI3K and Akt1. Carcinogenesis *31*, 1604–1611.

Smith, D.R., Myint, T., and Goh, H.S. (1993). Over-expression of the c-myc proto-oncogene in colorectal carcinoma. Br. J. Cancer *68*, 407–413.

Spano, J.P., Fagard, R., Soria, J.-C., Rixe, O., Khayat, D., and Milano, G. (2005). Epidermal growth factor receptor signaling in colorectal cancer: preclinical data and therapeutic perspectives. Ann. Oncol. Off. J. Eur. Soc. Med. Oncol. *16*, 189–194.

Szklarczyk, D., Franceschini, A., Wyder, S., Forslund, K., Heller, D., Huerta-Cepas, J., Simonovic, M., Roth, A., Santos, A., Tsafou, K.P., *et al*. (2015). STRING v10: protein-protein interaction networks, integrated over the tree of life. Nucleic Acids Res. *43*, D447-452.

Tompa, P. (2003). Intrinsically unstructured proteins evolve by repeat expansion. BioEssays *25*, 847–855.

Uversky, V.N., Davé, V., Iakoucheva, L.M., Malaney, P., Metallo, S.J., Pathak, R.R., and Joerger, A.C. (2014). Pathological Unfoldomics of Uncontrolled Chaos: Intrinsically Disordered Proteins and Human Diseases. Chem. Rev. *114*, 6844–6879.

Wu, S., Wan, P., Li, J., Li, D., Zhu, Y., and He, F. (2006). Multi-modality of pI distribution in whole proteome. PROTEOMICS *6*, 449–455.

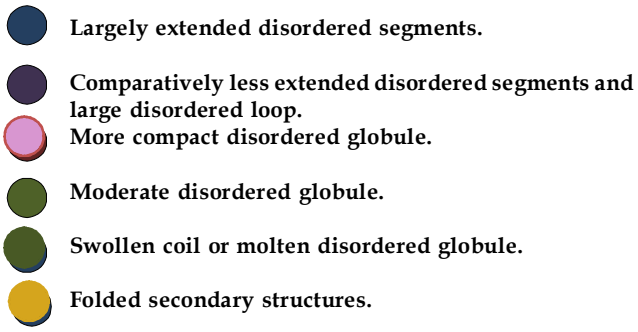
Yang, X.-J. (2004). The diverse superfamily of lysine acetyltransferases and their roles in leukemia and other diseases. Nucleic Acids Res. *32*, 959–976.

Yang, C., Xie, X., Tang, H., Dong, X., Zhang, X., and Huang, F. (2018). Transcriptome analysis reveals GA induced apoptosis in HCT116 human colon cancer cells through calcium and p53 signal pathways. RSC Adv. *8*, 12449–12458.

Yao, Y.-L., Shao, J., Zhang, C., Wu, J.-H., Zhang, Q.-H., Wang, J.-J., and Zhu, W. (2013). Proliferation of Colorectal Cancer Is Promoted by Two Signaling Transduction Expression Patterns: ErbB2/ErbB3/AKT and MET/ErbB3/MAPK. PLOS ONE *8*, e78086.

Zhu, N., Qin, L., Luo, Z., Guo, Q., Yang, L., and Liao, D. (2014). Challenging role of Wnt5a and its signaling pathway in cancer metastasis (Review). Exp. Ther. Med. *8*, 3–8.

**Color Key-**



**Largely extended disordered segments.**

**Comparatively less extended disordered segments and large disordered loop.**

**More compact disordered globule.**

**Moderate disordered globule.**

**Swollen coil or molten disordered globule.**

**Folded secondary structures.**