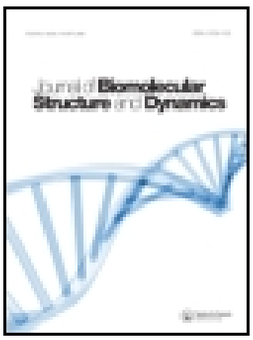


Taylor & Francis
Taylor & Francis Group

Journal of Biomolecular Structure and Dynamics

ISSN: 0739-1102 (Print) 1538-0254 (Online) Journal homepage: http://www.tandfonline.com/loi/tbsd20

Structural Disorder: A tool for housekeeping proteins performing tissue-specific interactions

Sanghita Banerjee & Rajat K. De

To cite this article: Sanghita Banerjee & Rajat K. De (2015): Structural Disorder: A tool for housekeeping proteins performing tissue-specific interactions, Journal of Biomolecular Structure and Dynamics, DOI: 10.1080/07391102.2015.1095115

To link to this article: http://dx.doi.org/10.1080/07391102.2015.1095115

Publisher: Taylor & Francis

Journal: *Journal of Biomolecular Structure and Dynamics*

DOI: http://dx.doi.org/10.1080/07391102.2015.1095115

Structural Disorder: A tool for housekeeping proteins performing tissue-specific interactions

Sanghita Banerjee*, Rajat K. De

Machine Intelligence Unit, Indian Statistical Institute, 203 Barrackpore Trunk Road, Kolkata 700108, India

SB: (banerjee.sanghita@gmail.com)
RKD: (rajat@isical.ac.in)

* Corresponding Author: banerjee.sanghita@gmail.com
FAX: +91-33-25783357
Telephone: +91-33-25753105/3100 (Laboratory)

Short Title: Tissue-specific interactions of housekeeping proteins.

Abstract

An interaction between a pair of proteins unique for a particular tissue is denoted as a tissue-specific interaction (TSI). Tissue-specific (TS) proteins always perform TSIs with a limited number of interacting partners. However, it has been claimed that housekeeping (HK) proteins frequently take part in TSIs. This is actually an unusual phenomenon. How a single HK protein mediates TSIs, - remains an interesting yet an unsolved question. We have hypothesized that HK proteins have attained a high degree of structural flexibility to modulate TSIs efficiently. We have observed that HK proteins are selected to be intrinsically disordered compared to TS proteins. Therefore, the purposeful adaptation of structural disorder brings out special advantages for HK proteins compared to TS proteins. We have demonstrated that TSIs may play vital roles in shaping the molecular adaptation of disordered regions within HK proteins. We also have noticed that HK proteins, mediating a huge number of TSIs, have a greater portion of their interacting interfaces overlapped with the adjacent disordered segment. Moreover, these HK proteins, mediating TSIs, preferably adapt single domain. We have concluded that HK proteins adapt a high degree of structural flexibility to mediate TSIs. Besides, having a single domain along with structural flexibility is more economic than maintaining multiple domains with a rigid

structure. This assists them in attaining various structural conformations upon binding to their partners, thereby designing an economically optimum molecular system.

Keywords: Intrinsically disordered proteins; unstructured proteins; protein domains; single domain proteins; TSI index.

Abbreviations

HK = housekeeping, TS = tissue-specific, IDP = intrinsically disordered protein, TSI = tissue-specific interaction, MD = multi-domain protein, SD = single domain protein.

1. Introduction

Genes and their products are classified as housekeeping and tissue-specific depending on the nature of their expression pattern. Housekeeping (HK) genes constitutively express in almost every tissue under normal conditions and maintain essential cellular functions (Eisenberg and Levanon 2003; Eisenberg and Levanon 2013). In contrast, tissue-specific (TS) genes express predominantly in a selective number of tissues and are responsible for particular tissue-specific functions. Studies, associated with genomic and proteomic features (sequence conservation rate (Eisenberg and Levanon 2003), evolutionary rate (Zhang and Li 2004), length of introns (Farré et al. 2007), untranslated regions and coding sequences, gene compactness , gene expression (Chang et al. 2011), and protein-protein interactions (Bossi and Lehner 2009)), so far, have shown a sharp discrimination between the groups of HK and TS genes. We have highlighted that

the effects of differential gene expression pattern between the groups of genes encoding HK and TS proteins eventually reflect upon the behavior of their protein-protein interactions.

Interestingly, Bossi et al. have claimed that most of the HK proteins take part in tissue-specific interactions (Bossi and Lehner 2009). TS proteins will obviously mediate TSIs, since their expression is limited to a particular tissue. However, a HK protein undergoing TSIs is an unusual phenomenon. A specific HK protein interacts with TS proteins in a particular tissue (where both the HK and TS proteins are co-expressed); making the interaction unique for that particular tissue, is termed as a tissue-specific interaction (TSI) (Figure S1). Therefore, we have assumed that HK proteins, being ubiquitously expressed in all the tissues, do participate in a large number of TSIs. However, the interacting partners of HK protein vary from one tissue to another. It gives rise to the questions like, how a single HK protein evolves to interact with a number of structurally diversified molecular targets under varied tissue environments. Does the class of HK proteins acquire any kind of specialization in their pattern of protein folding to regulate TSIs, in the course of evolutionary time? Structural conformations and folding pattern of a particular protein assist them in binding to their counter interacting partners. Hence, it has been believed that the complete comprehension of the structural integrity and dynamics of the set of proteins will gradually resolve the entire set of protein interactions or the interactome.

To date, a few attempts have been made to study the variation in structural features of proteins encoded by the classes of HK and TS genes. In this scenario, our study has demonstrated that HK proteins tend to be intrinsically more disordered compared to TS proteins. The stretches of disordered regions within HK proteins are highly conserved across the mammalian species, whereas, in the case of TS proteins, they evolve rapidly. Towards this goal, we have hypothesized that HK proteins adapt a high degree of plasticity of structural conformations to

undergo TSIs. Therefore, in our present study, we have investigated how do HK proteins exploit the presence of intrinsic disorder to modulate TSIs? The question becomes interesting as we have studied further.

In the context, some recent studies have mentioned several advantages of intrinsically disordered proteins (IDPs) undergoing efficient protein interactions. The study of Pfam domains has revealed that many interacting domains are intrinsically disordered and evolutionarily conserved in nature (Tompa et al. 2009). These disordered domains usually behave as the molecular recognition sites (Mészáros et al. 2007; Gruet et al. 2013). On the other hand, in many cases, disordered segments act as flexible linkers between different subunits of a multi-domain protein. It is believed that the structural flexibility has enhanced their adaptability of binding to various targets (Gunasekaran et al. 2003). Intrinsically disordered regions can undergo temporary conversion to required structured forms upon binding to molecular targets (Receveur-Bréchot et al. 2006; Boehr et al. 2009; Wright and Dyson 2009). On the other hand, the extreme flexibility of disordered regions sometimes becomes disadvantageous for the cellular systems. It causes IDPs to get involved in unwanted interactions at the high concentration (Babu et al. 2011). Thus, the over-expression of such proteins is detrimental for the cellular systems (Vavouri et al. 2009). Therefore, the regulation of IDPs is highly controlled in various stages of the central dogma (Gsponer et al. 2008). In this background, it is intriguing to explore how the trend of evolution has exploited the existence of structural disorder balancing their pros and cons.

In this study, we have reported a difference in the distribution of various parameters (i.e., overall disordered residues, number of disordered regions and length of disordered regions) measuring the extent of structural disorder within the groups of HK proteins mediating varying degree of TSIs. The conditional probability test has indicated that TSIs actually shapes the adaptation of

disordered regions within human HK proteins. We have also noticed that the adaptation of a few long disordered segments is comparatively more favorable for modulating TSIs than multiple small disordered regions. In order to support this result, we have studied further on the types of protein domains adapted by the classes of HK proteins undergoing a huge number of TSIs. We have come up with the conclusion that HK proteins modulating TSIs prefer to maintain single domains along with a high degree of structural flexibility rather than adapting multiple domains with rigidity in protein folding. These instances unfold how the the evolution of intrinsically disordered regions balances itself to combat the necessity of cellular systems.

2 Materials and Methods

In this section, we have described the detailed methods of identifying HK and TS genes using RNA sequencing data. We have mentioned the step-wise procedure for the prediction of human TSIs integrating gene expression and protein-protein interaction datasets (Figure S2). Additionally, we have also included the methods for identifying intrinsically disordered residues, protein domains, and their expression.

2.1 Identification of housekeeping and tissue-specific proteins using RNA sequencing data

We have preferred using RNA sequencing datasets for measuring gene expression (Grabherr et al. 2011; Zhao et al. 2011). Therefore, we have considered the expression data given in terms of RPKM (Reads Per Kilobase of exon per Million reads mapped) values for each gene in 11 normal human tissue types provided by RNA-Seq Atlas (RSA) (http://medicalgenomics.org/rna_seq_atlas/download) (Grabherr et al. 2011), for identifying the sets of HK and TS genes. We have extracted the housekeeping transcripts following the criteria given by the recent study of Eisenberg et al (Eisenberg and Levanon 2013). According to the

criteria, a housekeeping transcript: i) must be expressed in all the tissues (i.e., RPKM value ≠ 0), ii) should have minimum variance in expression values across all the tissues (i.e., standard deviation of RPKM value < 1), and iii) should not exhibit abrupt fluctuation (either higher or lower) of expression value in a single tissue (i.e., the RPKM value in each tissue should not differ from the average RPKM value across all the tissues by 2 or 4 folds).

In our method, we have purposely considered the expression profile of transcripts (and not genes) to avoid the issues arising due to the phenomenon of alternative splicing. However, we have considered the genes corresponding to the housekeeping transcripts for further study. The above RSA data has provided a set of 6977 HK genes which are mostly found common with the set of HK genes given by Eisenberg et al. (http://www.tau.ac.il/~elieis/HKG/) (Eisenberg and Levanon 2013). We have followed the method of Eisenberg et al (Eisenberg and Levanon 2013). They have first used the criterion of low expression variation for defining HK genes that was ignored in earlier microarray studies.

On the other hand, we have identified the TS transcripts from RNA sequencing expression data using the criteria: i) RPKM value should be non-zero at least in one tissue; ii) RPKM values should have high variance across the tissues; and iii) RPKM values of a transcript in at most two tissues must be four/eight folds higher compared to that in other tissues, (i.e., the RPKM values in each of these tissues should be greater than 4 or 8 folds than the average RPKM value across all these 11 tissues). Additionally, we have worked on the data provided by Human Protein Atlas (HPA) (http://www.proteinatlas.org) (Uhlén et al. 2015) that comprises FPKM (Fragments Per Kilobase of exon per Million reads mapped) values of 20314 transcripts of human protein-coding genes across 44 cell lines and 27 tissues, obtained by RNA sequencing method. We have

considered the expression values (FPKM) across 44 cell lines and 27 tissues separately, and used them further for identifying the sets of HK and TS genes based on the criteria mentioned above.

We have primarily tested the hypothesis using RNA-Seq Atlas (RSA) (http://medicalgenomics.org/rna_seq_atlas/download). However, in order to obtain an unbiased result, we have also considered the gene expression data generated by another RNA Sequencing dataset (Human Protein Atlas – http://www.proteinatlas.org/). We have obtained the sets of HK and TS proteins from Human Protein Atlas following a similar method applied to RNA-Seq Atlas.

2.2 Identification of tissue-specific interactions

We have retrieved a large-scale protein-protein interaction data, compiled by Bossi et al. using 21 databases of human interactome (Bossi and Lehner 2009). We have integrated the information on the pairwise protein-protein interactions with the gene expression datasets (i.e., RSA and HPA datasets). We have retained only those interactions in which both the interacting proteins have the expression values available in the dataset (Figure S2).

We have used the above data to retrieve the "tissue-specific" interactions (TSIs). We have called an interaction "tissue-specific" if the pair of interacting proteins co-expresses in at most two tissues. In order to avoid any ambiguity, it is to be mentioned here that if the interacting pair of proteins co-expresses in more than two tissues, the interaction will no longer be considered as a TSI. We have counted the number of TSIs carried out by each protein, and denoted it as the tissue-specific interaction index or "TSI index" of that particular protein. It is expected that a TS protein will only perform TSIs, as the corresponding gene expression is limited to a few tissues.

Strikingly, we have found that some of the HK proteins frequently undergo TSIs, although the number of TSI, mediated by HK proteins is comparatively smaller than those of TS proteins. Thus, there appears an incomparable difference in the distribution of the TSI index values of HK and TS proteins. Our observation has also been supported by Bossi et al. (Bossi and Lehner 2009). Hence, we have set the threshold values of TSI index separately for the sets of HK and TS proteins. We have primarily conducted our study categorizing the set of HK proteins into three groups based on their average TSI index (mentioned in Section 3.3). However, we have also worked with flexible cut-offs to maintain the robustness of our results (mentioned in Section 3.3).

2.3 Identification of intrinsically disordered and structured proteins

We have identified the disordered or ordered state of every residue within the human and mouse proteome using IUPred-Long (IUPred-L) (Dosztányi et al. 2005), ESpritz (Walsh et al. 2012), and PONDR-FIT servers (Xue et al. 2010). We have primarily used the measure given by the IUPred-L server (http://iupred.enzim.hu/). IUPred-L uses a sequence based algorithm for measuring the ability of amino acid residues to form stabilizing contacts. The method assumes that amino acid residues having less inter-residue interactions are likely to be more disordered in nature due to lack of stabilizing energy, usually required during protein folding. The L version of the IUPred method denotes that the algorithm particularly runs on the long stretches of disordered regions. According to the scores provided by the method of IUPred-L, an amino acid residue is considered to be disordered if its value is greater than 0.5 and ordered if it is at most 0.5 (Dosztányi et al. 2005). On the basis of the predicted disordered residues returned by the IUPred-L tool, we have calculated several parameters, like i) number of disordered residues (DC), ii) number of disordered regions (DR count; each regions having ≥ 30 consecutive

disordered residues), and iii) length of the disordered regions (DL) within an entire protein sequence. As the parameter values depend on the length of the entire protein, we have normalized them with the sequence length. Hence, the values of the parameters will be independent of the entire protein length. We have used this information to categorize the sets of HK and TS proteins into different groups viz., i) highly disordered or unstructured (IDPs), if DC > 30 and contains at least one disordered region within the protein sequence, ii) moderately disordered (M-IDPs), where 10 < DC ≤ 30, and iii) ordered or well-structured (STRs), if DC ≤ 10. The above classification has been done following the study of Gsponer et al (Gsponer et al. 2008). In order to maintain the stringency, we have used only the extreme groups – IDPs and STRs and have ignored the group M-IDPs. We have measured both independent and cumulative effects of all the three parameters (i. DC, ii. DR count, and iii. DL) on the extent of TSIs mediated by a protein (Section 3.3). We have further validated the result using other disorder prediction tools (PONDR-FIT and ESpritz) to ensure that the result is independent of the prediction methods used to measure the disordered residues (Section 3.3). PONDR-FIT uses an artificial neural network prediction method, which was designed as a meta-predictor combining the outcomes of several individual disorder predictors (Xue et al. 2010; Habchi et al. 2014). On the other hand, ESpritz predicts structural disorder using bi-directional recursive neural networks (Walsh et al. 2012).

2.4 Determination of the evolutionary rate of disordered regions

The ratio of substitution occurring at non-synonymous (d_N) and synonymous (d_S) sites, is used as an effective measure in evolutionary studies to calculate the extent to which selection pressure acts on gene sequence evolution. Non-synonymous (d_N) substitution causes divergence with respect to amino-acid content while the synonymous (d_S) substitution does not alter the amino

acid content of the sequence. We have calculated the evolutionary rates (d_N/d_S ratio) comparing human *(Homo sapiens)* gene sequences against one-to-one orthologous sequences from those of the mouse *(Mus musculus)*. We have obtained the entire coding sequences of both human and mouse genomes from the NCBI RefSeq server (Pruitt et al. 2005). We have used EMBOSS Needle algorithm (McWilliam et al. 2013) in order to obtain the pair-wise alignment for each set of orthologous gene pairs. It applies Needleman-Wunsch alignment algorithm to find the optimum alignment (including gaps) of two sequences along their entire length. The evolutionary rates (d_N/d_S) have been calculated from the pair-wise alignment using Yang and Nielsen method (Yang and Nielsen 2000) given in the PAML package (version 4). Furthermore, we have separated the disordered and structured regions from an IDP within the pair-wise sequence alignment, using a Perl script, and have calculated the evolutionary rates of disordered and structured region individually, in the similar way as mentioned above.

2.5 Identification of protein domain interfaces and Molecular Recognition Elememts (MoREs)

We have retrieved the set of protein domains from the Pfam repository (http://pfam.sanger.ac.uk/) (Punta et al. 2012). We have mapped HK and TS proteins bearing the protein domains, and calculated the number of both unique and repetitive domains for each protein. Following the study of Kim et al. (Kim et al. 2008), we have determined the cut-off value for assigning protein domain, as (1) e-value of alignment < 1×10^{-4}, (2) overlapped sequence length > 80% of domain length and (3) domain length > 10 residues. According to the definition given by Kim et al. (Kim et al. 2006a), a protein having at most two domains is considered as a "single domain" (SD) protein, while that with many (> 2) domains is considered to be a "multi-domain" (MD) protein.

Furthermore, we have calculated certain parameters, like length and location of a domain region within a protein sequence. We have mapped the domain regions with the stretches of disordered regions in a protein and measured the fraction of disordered regions overlapped with the adjacent domain regions.

We have identified the Molecular Recognition Elememts (MoREs) using the prediction servers, like ANCHOR (http://anchor.enzim.hu/index_multi.php) (Dosztányi et al. 2009) and MoRFPred (http://biomine-ws.ece.ualberta.ca/MoRFpred/index.html) (Disfani et al. 2012). ANCHOR method predicts protein–protein binding residues located in the disordered regions. It implies the criteria to identify the disordered binding regions as: i) the stretch of residues (10-70) embedded within a long disordered region, ii) the residues should not form favorable contacts with neighboring residues (so lack of folding persists), iii) it should form favorable contacts with other globular proteins and iv) the residues gain stabilizing energy interacting with other globular proteins. On the other hand, MoRFPred finds molecular recognition elements (up to 25 consecutive residues) within the stretches of disordered segments using Support Vector Machine (SVM) classifier.

2.6 Tissue specificity of protein domains

We have further identified the protein domains, previously retrieved from Pfam database as "housekeeping" and "tissue-specific" depending on the expression pattern of the genes encoding these proteins. Lehner and Fraser have ranked the protein domains according to its tissue specificity (Lehner and Fraser 2004). They have used microarray dataset of GNF Gene Expression Atlas for measuring the expression level of ~10,000 mouse genes across 45 tissues. However, in our study, we have used RNA-Seq data to categorize the sets of protein domains as

"housekeeping" and "tissue-specific" ones. Accordingly, we have labeled a protein domain as "housekeeping" if its corresponding gene gets expressed in all the tissues, and as "tissue-specific" if it is over-expressed by 4 fold in at most two tissues or cell lines compared to the remaining tissues or cell lines.

2.7 Statistical Analysis

All the statistical tests and graphical plots of the data have been conducted using SPSS version 20.0 and R Statistical Package. As most of the datasets considered in our study are not normally distributed, we have exclusively applied non-parametric statistics that do not make any prior assumptions regarding the probability distribution of the variables being measured. We have thus purposely used the Mann-Whitney U test, a non-parametric statistical test, which assesses whether two samples are likely to come from the same underlying populations. The test can also be used to estimate whether the medians of two distributions are significantly different. Box-plots have been used as a non-parametric measure to graphically illustrate various statistical properties of the distribution, like median, variation in the given distributions, minimum and maximum values, interquartile range, outliers, along with the skewness. In order to maintain the robustness of the statistics, we have preferably used median values to represent the center of non-normally distributed data, instead of mean values. We have used Spearman rank correlation coefficient as a non-parametric measure for estimating the association (correlation) between two ranked variables.

3. Results

In this section, we have explored the mode of adaptation of structural disorder within the classes of HK and TS proteins, and have delved into the probable reasons behind such type of molecular

adaptation. For this purpose, we have analyzed the *in silico* results followed by making an appropriate hypothesis. We have performed the analyses using RNA sequencing datasets (RSA and HPA-tissues) to improve the accuracy of the results. In the following sections, we have stated only the results obtained from RSA, while we have reported the results obtained from the other dataset (HPA-tissues) in the Supplementary file as "Extended results using HPA dataset". (Figures S3–S5).

3.1 Structural disorder in housekeeping proteins

In order to access how the classes of human HK and TS proteins exhibit difference in the pattern of protein folding or unfolding, we have considered three parameters: i) number of disordered residues (DC) ii) number of disordered regions (DR count), and iii) length of the disordered regions (DL) within a protein, as the measures to estimate the extent of intrinsic disorder within the classes of human HK and TS proteins. We have observed that all these three measures of structural disorder are relatively higher within the class of HK proteins than those of TS proteins (Mann-Whitney U test, $P = 1.0\times10^{-6}$; Figures 1). We have repeated the analyses using other disorder predictor tools (ESpritz and PONDR-FIT) and come up with a similar trend of result (Figure S6).

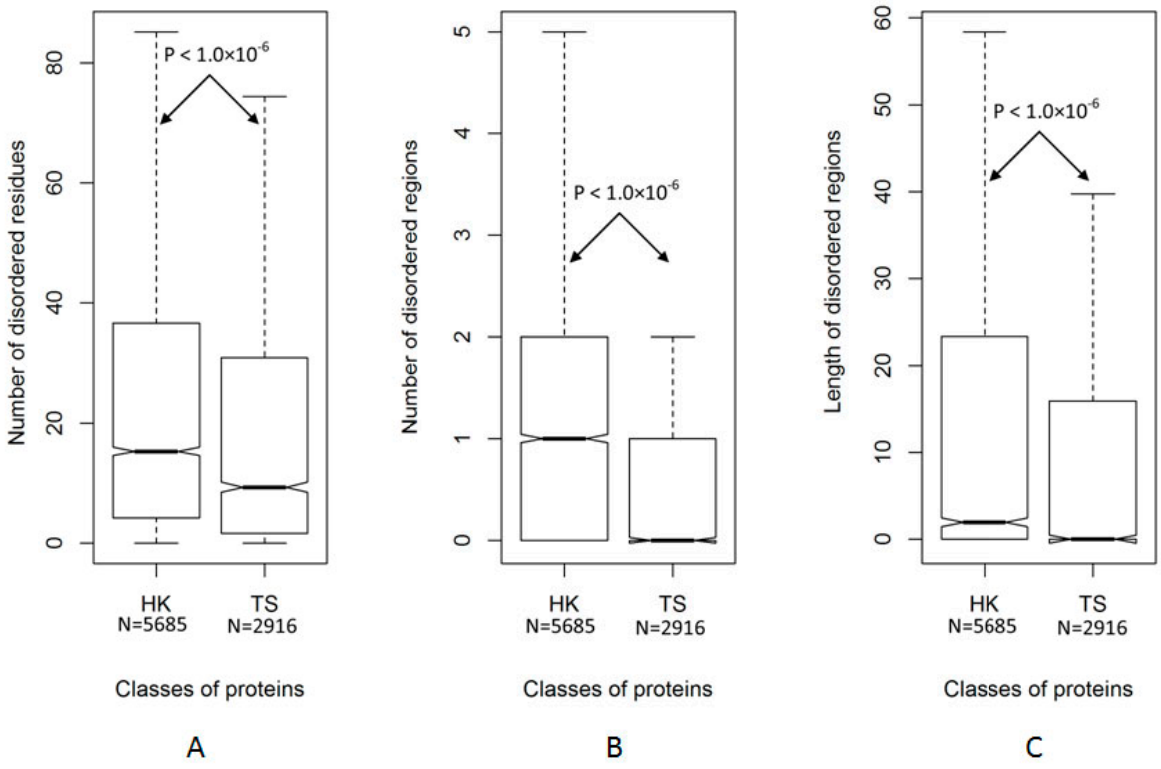


Figure 1: Difference in the measure of structural disorder between housekeeping and tissue-specific proteins. Boxplots showing the distribution of the three parameters - A) number of disordered residues, B) number of disordered regions, and C) length of disordered regions, measuring the extent of structural disorder between the classes of housekeeping (HK) and tissue-specific (TS) proteins.

Furthermore, to estimate the population of disordered HK proteins, we have classified the whole human proteome into three categories: i) highly disordered proteins (IDPs), ii) moderately disordered proteins (M-IDPs), and iii) well-structured proteins (STRs) (details in Section 2.3). Interestingly, we have observed that the pool of IDPs tends to behave more like an HK protein (Figure S7). In contrary, a few HK proteins are highly disordered in nature (31% HK and 69% TS proteins within the entire set of IDPs). It indicates that only a small subset of HK proteins is

highly enriched in structural disorder, which in turn, increases the overall measure of the structural disorder within the entire set of human HK proteins when compared with the set of TS proteins (Figures 1).

Further, we have noticed that the small set of disordered HK proteins is not only structurally disordered in nature, but also maintains the characteristics of a typical IDP. We know that gene expression breadth (EB) correlates positively and strongly with the expression level (Park and Choi 2010). However, in contrary, our result suggests that intrinsically disordered HK proteins exhibit significantly lower expression level compared to those of other well-structured HK proteins, in spite of having almost similar EB (Figure S8)**.** The result is in agreement with the existing literature which claims that high expression of disordered proteins comes with a high fitness cost (Tomala and Korona 2013). Taken together, these results suggest that the adaptation of structural disorder within a small subset of HK proteins is designed purposefully to render some special cellular functions.

3.2 Evolutionary conservation of disordered regions within housekeeping and tissue-specific proteins

We have focused on the evolutionary variation of disordered regions in human (*Homo sapiens*) from those of the mouse *(Mus musculus),* with an aim to capture their evolutionary importance within HK and TS proteins. According to the literature, intrinsically disordered regions are frequently subjected to a high rate of substitutions due to the lack of structural constraints (Brown et al. 2002; Brown et al. 2011). However, the evolutionary pressure (d_N/d_S) (Mann-Whitney U test: $P=1.67\times10^{-13}$, Figure S9A) and non-synonymous substitutions (d_N) (Mann-Whitney U test: $P=1.0\times10^{-8}$, Figure S9B) acting on the stretches of intrinsically disordered

regions indicates that the disordered regions within TS proteins undergo a high rate of substitution while those regions in HK proteins prefer to be rather conserved. Besides, the difference in the distribution of synonymous substitution (d_S) between HK and TS proteins is weak (Mann-Whitney U test: $P=6.7\times10^{-5}$, Figure S9C).

These observations suggest that there might be some evolutionary advantages of maintaining the stretches of disordered regions. However, at the same time, proteins with long disordered stretches are found to have a high tendency to get into misinteractions while in abundance (Vavouri et al. 2009). This posed to be a major disadvantage for the cellular system. Moreover, it is believed that HK proteins handle the basic and some of the essential functions of the cellular system. In this scenario, we have raised a vital question: how does the cellular system balance between the beneficial and detrimental effects of disordered stretches? Consequently, we have hypothesized that the molecular adaptation of disordered regions within human HK proteins, perhaps renders some special functional efficiencies (benefits) which actually justify the cellular system taking the risk (costs). On this ground, the adaptation of structural disorder in HK proteins, over rigid protein folding, is favorable to optimize the energy resource of the biological systems. In order to understand - how the lack of a well-defined structure in HK proteins benefits the cellular system, we have extended our results as follows.

3.3 Enrichment of structural disorder in housekeeping proteins mediating tissue-specific interactions

Experimental investigations have already focused that the property of molecular interactions adapted by a particular protein is greatly facilitated through its structural orientations (Kim et al. 2008; Zhang et al. 2012). We have concentrated on the type of protein-protein interactions

mediated by HK proteins as a probable factor influencing the adaptation of structural disorder within HK proteins. To this end, we have considered the phenomenon of tissue-specific interactions (TSIs). TS proteins only perform TSIs. However, TSI mediated by HK protein is indeed an interesting phenomenon (Figure S1). Many HK proteins are found interacting with HK or non-HK proteins (i.e., either TS or other proteins) in a particular tissue, making the interaction unique for the tissue. It delineates the tissue-specific role of a HK protein (Figure 2). Genes, encoding HK proteins, are expressed ubiquitously across the tissues, but their mode of interactions with various interacting partners varies across the tissues. In contrast, a TS protein interacts with a limited number of proteins (HK or non-HK), and they are co-expressed only in that particular tissue. It was previously thought that TSIs are partially regulated by monitoring the level of gene expression of HK proteins (Warrington et al. 2000; Briggs et al. 2011). However, to date it has not been studied extensively that how a human HK protein manages itself to bind to a number of structurally different partners under different conditions in various tissues. It questions if there is any special structural feature that benefits them in bringing about TSIs, efficiently.

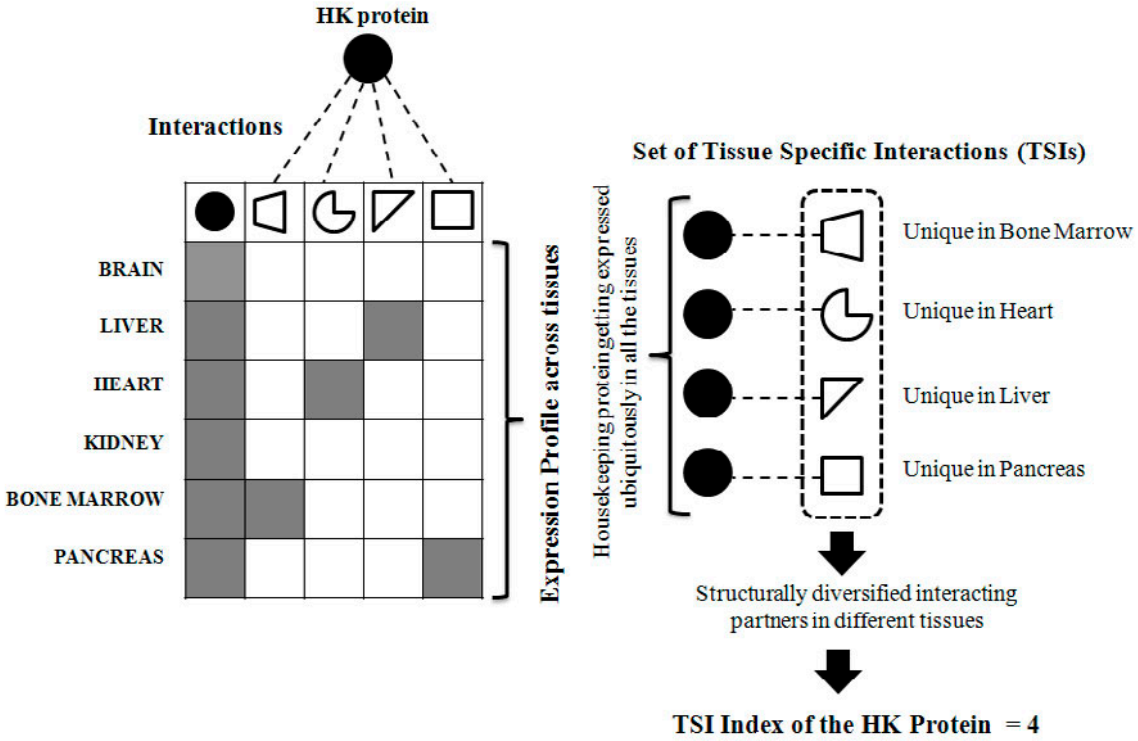


Figure 2: Mechanism of tissue-specific interactions mediated by housekeeping proteins. Schematic diagram explains how the increase in tissue-specific interaction index (TSI index) of a particular HK protein, in turn, increases the variation and the number of interacting partners.

In order to answer these questions, we have explored if there is any interrelation between structural disorder and TSIs within the groups of human HK and TS proteins. We have found that TS proteins, as expected, have a very high TSI index value compared to that of HK proteins (P=1.0×10^{-6} significant according to Mann-Whitney U test, Figure S10). Hence, it will cause a huge disparity in the analysis of the results. Therefore, we have considered only the class of HK proteins for further analysis. Accordingly, we have categorized the set of HK proteins, having TSIs depending on their average TSI index values (≈ 2.9), into i) P_{HTSI} (TSI index value = 3 to 81), ii) P_{LTSI} (TSI index value = 1 to 2), and iii) P_{NOTSI} (TSI index value = 0). Then, we have estimated all the three features (i.e., 1. Number of disordered residues (DC) 2. Number of disordered regions (DR count) and 3. Length of the disordered regions (DL)), which are

considered as the measures of proteins' structural flexibility and stability, within the three groups of $P_{HTSI,}$ P_{LTSI} and P_{NOTSI}. We have observed a difference in the distribution of all these three parameters between the three groups (Figure 3). The results claim that P_{HTSI} are enriched in disordered residues compared to the other two groups of HK proteins participating in less (P_{LTSI}) or no TSI (P_{NOTSI}) (P=1.0×10^{-10} significant according to Mann-Whitney U test, Figure 3A). Additionally, we have noticed that there is hardly any difference in the number of disordered regions within the groups of P_{HTSI} and P_{LTSI}, while both these groups differ significantly from the group of HK proteins not undergoing TSI (P_{NOTSI}) (P=3.3×10^{-17} using Mann-Whitney U test, Figure 3B). However, P_{HTSI} tends to have longer disordered regions compared to the groups of P_{LTSI} and P_{NOTSI} (P=1.0×10^{-10} using Mann-Whitney U test, Figure 3C). Together, these findings indicate that the set of HK proteins participating in a higher number of TSIs, favors the presence of the long stretches of disordered segments.

We have further assured that the above conclusion remains unaffected by the use of various disorder prediction algorithms (Figure S11). Moreover, we have repeated the analysis using flexible cut-offs of TSI index values, to test the robustness of the result obtained from the categorization of HK proteins into P_{HTSI}, P_{LTSI} and P_{NOTSI}. The outcome suggests a similar trend of the result (detailed results in Table S1, Figures S12 – S16) indicating that the observation is independent of selecting TSI index threshold values.

On the other hand, there is no significant difference in the distribution of the total disorder content (DC) and the number of disordered regions (DR count) between the groups of P_{HTSI} and P_{LTSI} within TS proteins. In this case, we have not considered the group P_{NOTSI} since all the TS proteins in our dataset have at least one TSI. We have strikingly noticed that P_{HTSI} contains relatively more number of disordered regions compared to P_{LTSI} in the group of TS proteins (P =

3.3×10^{-17} using Mann-Whitney U test), which is in contrary to the result obtained from the set of HK proteins.

Disordered proteins are found to have a high degree of interactions compared to those of well-structured ones (Haynes et al. 2006). We have also noticed that the number of protein-protein interactions strongly correlates with TSI index (Spearman $\rho = 0.74$, significant at $P = 1.0 \times 10^{-4}$). In this context, one can argue that the higher enrichment of intrinsic disorder within P_{HTSI} might be an artifact of the higher degree of protein-protein interactions mediated by HK proteins. However, the partial correlation test has rejected this probability, since the overall disorder content (r = 0.094, controlling factor = number of protein-protein interactions; $P = 1.0\times10^{-11}$) significantly correlates with the TSI index of a HK protein, while no such significant correlation has turned up in case of protein-protein interactions (r = 0.021, controlling factor = TSI index; not significant). Thus, the result illustrates that the association between the structural disorder and the TSI index of a HK protein is independent of the degree of protein-protein interactions. Hence, the above result is in agreement with our proposed hypothesis.

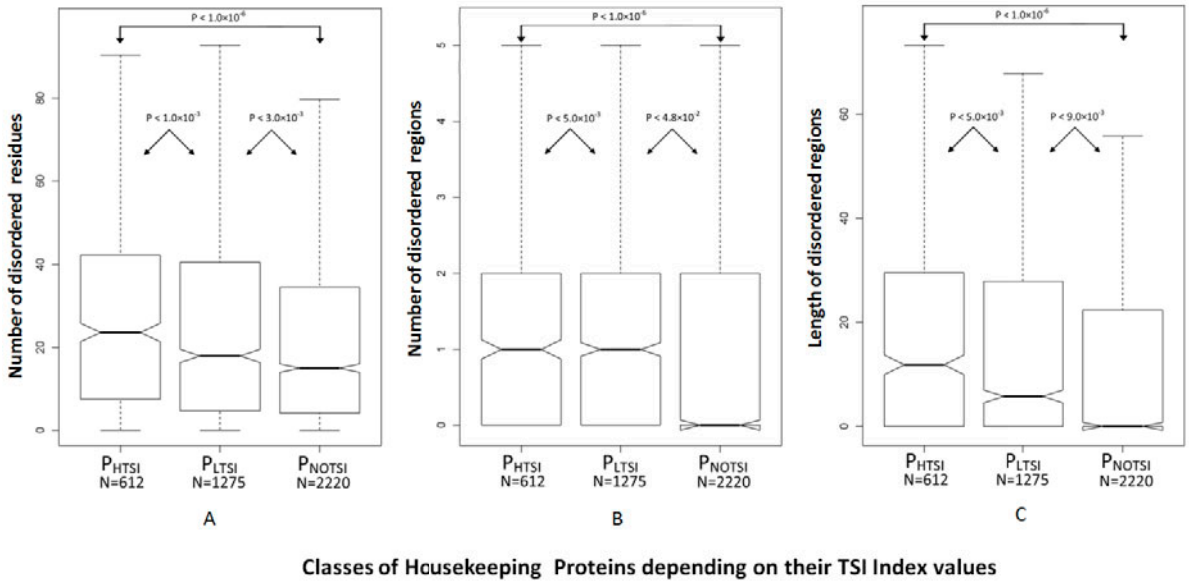


Figure 3: Distribution of structural disorder among the groups of housekeeping proteins having a varying degree of TSI index. Box-plots showing the distribution of three parameters [A: percentage of disordered residues within the entire protein, B: number of disordered regions and C: length of disordered regions) measuring structural disorder within the classes of housekeeping (HK) proteins having high TSI index (P_{HTSI}), low TSI index (P_{LTSI}) and those not undergoing any TSI (P_{NOTSI}).

3.3.1 Tissue-specific interactions favor the adaptation of a few, but long disordered regions

The correlation test indicates that TSI index exhibits positive, albeit weak, correlation separately with i) DC (Spearman ρ = 0.087, P=1.1×10^{-89}), and ii) DL (Spearman ρ = 0.076, P = 1.0×10^{-6}). However, the correlation coefficient value decreases between TSI index and DR count (Spearman ρ = 0.013, P=4.9×10^{-16}). We have suggested that among the three parameters of structural disorder, the influence of DC and DL is likely to be more than the DR count. To analyze further, we have clustered the set of disordered proteins, based on the length of the disordered regions (DL), as 1.Very long (DL > 60), 2.Long (40 > DL ≤ 60), 3.Medium (20 > DL ≤ 40) and 4.Small (DL = 1 to 20), and have found that the proteins having extremely large disordered regions are prone to get into a maximum number of TSIs (Figure S17A). However, the difference in the distribution is not statistically significant. The results may suggest one of the two possible conclusions - whether the TSI index influences: i) the presence of a very few, but long disordered regions, or ii) multiple small disordered motifs. Note that, in both the cases, the difference will be reflected in the length of disordered regions, since we have calculated the DL values as the sum of all disordered regions.

To investigate the issue, we have further categorized the entire set of HK proteins into three groups based on the numbers of disordered regions (DR): (i) well-structured proteins (DR count = 0), (ii) disordered proteins having a single disordered region (DR count = 1), and (iii) other

disordered proteins having multiple disordered regions (DR count = 2 to 23). We have observed that the set of disordered proteins having a single disordered region has the highest TSI index value compared to the other groups (Figure S17B). Additionally, the proteins having a single and long disordered region, preferably undergo a higher number of TSIs compared to those having a single, but short disordered region (Figure S17C). It thus suggests that the natural selection favors the molecular adaptation of a single (a few) long disordered region(s), instead of multiple short ones for the purpose of TSIs.

3.3.2 Tissue-specific interactions and structural disorder – which one is the cause and which one is the effect?

Hitherto, we have predicted an association between the structural disorder and TSIs in case of human HK proteins. However, the result is not appropriate to claim the actual cause and its effects, i.e., if structural disorder influence TSIs or the reverse trend of TSIs to drive the molecular adaptation of intrinsic disorder in HK proteins. For this purpose, we have performed the conditional probability test which establishes that HK proteins undergoing TSIs have a high probability of acquiring structural flexibility in them, but conversely, it is unlikely that the class of unstructured proteins will necessarily undergo TSIs (Case A, Table 1). However, a protein can acquire intrinsically disordered residues due to several biological reasons, beside assisting TSIs. Additionally, the results demonstrate a gradual increase in the number of disordered residues with the increase in the TSI index value (Case B, Table 1). We have thus suggested that the molecular adaptation of structural disorder within human HK proteins is purposefully designed to mediate TSIs efficiently and, therefore, the reverse condition may not be true.

Table 1: **Conditional probability test for structural disorder and tissue-specific interactions (TSIs).**

Protein Category	**Tissue-specific interactions (TSIs)**			
	High-TSI (TSI index = 3 to 81)	**Low-TSI (TSI index = 1, 2)**	**No-TSI (TSI index = 0)**	**Total**
STR	373	835	1550	2758
IDP	227	423	651	1301
Total	600	1258	2201	**4059**

Case	**Event (E)**	**Condition (C)**	**Probability (E|C) = P(E∩C)/P(C)** i.e., Conditional probability of (E|C) indicates that the probability of observing an event E given the condition C.

A	**High-TSI**	**IDP**	The probability of finding a protein with high TSI index among the set of intrinsically disordered proteins (IDPs): P (High-TSI | IDP) = $\frac{227}{1301} = 0.17$ Comparatively, the low probability value indicates intrinsically disordered proteins are unlikely to undergo tissue-specific interactions. But, the reverse condition might be true.
	IDP	**High-TSI**	The probability of finding a disordered protein (IDP) among the group of proteins having high TSI: P (IDP | High-TSI) = $\frac{227}{600} = 0.38$ The high probability value indicates that a HK protein undergoing tissue-specific interactions is more likely to be intrinsically disordered in nature, but the reverse is not true. A disordered protein (IDP) may not necessarily undergo tissue-specific interactions.
B	**IDP**	**High-TSI**	The probability of finding a disordered protein (IDP) among the group of proteins having a high TSI index: P (IDP | High-TSI) = $\frac{227}{600} = 0.38$ The high value suggests that HK proteins having a high TSI index influence the adaptation of structural disorder.
	IDP	**Low-TSI**	The probability of finding a disordered protein (IDP) among the group of proteins having a low TSI index: P (IDP | Low-TSI) = $\frac{423}{1258} = 0.33$ The lower value suggests that the urge of maintaining structural disorder within HK proteins having a low TSI index is comparatively lower than those HK proteins having a high TSI index.

	IDP	**No-TSI**	The probability of finding a disordered protein (IDP) among the group of proteins not undergoing TSI: P (IDP | No-TSI) = $\frac{651}{2201} = 0.29$ The probability indicating that HK proteins not mediating any TSI, will favor the adaptation of structural disorder, is minimal.

3.4 Structural adaptability of the disordered regions as molecular binding sites

A protein domain or active site is assumed to be an actual physical site that takes part in protein interactions. In order to assess the structural behavior of the domains within HK proteins, particularly undergoing TSIs, we have examined if protein domains are themselves disordered in nature or the stretches of disordered regions connect two well-structured domains in a protein. In this context, we have noticed that the fraction of disordered regions overlapping with a protein domain almost increases with the decrease in the unique domain number (UDN) within the set of HK proteins (Figure S18). However, the difference is not significant between the classes of HK proteins having a single domain (i.e., UDN = 1), and other individual groups having UDN = 4 to 7 and 14 due to incomparable sample sizes (N). The result suggests that proteins with a single domain (UDN = 1) tend to have a maximum portion of their domain overlapped with the adjacent disordered stretches.

Similarly, we have analyzed the existence of molecular recognition elements (MoREs) and a potential binding site (predicted using ANCHOR) within the disordered regions. Molecular recognition elements (MoREs) are short protein regions within longer disordered sequences that specifically participate in protein-protein interactions (Oldfield et al. 2005). MoREs allow a great extent of binding diversity of a region and thus recognizes differently structured molecular

partners at the binding site (Mohan et al. 2006; Baronti et al. 2015). We have found no such significant difference in the presence of both disordered binding sites and MoREs between the classes of HK and TS proteins (Figure S19 and S20), except in the number of disordered binding sites (Figure S19A). However, a significant increase in the existence of both disordered binding sites (Figure S19) and MoREs (Figure S20) turned up with the increase in the value of the TSI index within different groups of HK proteins having different degree of TSI index. Additionally, HK proteins having a high TSI index maintain a longer stretch of disordered binding segments compared to other groups. We have assumed from the result that the set of HK proteins may not necessarily have a disordered binding region. The requirement of having a MoRE or a potential binding region within the disordered segment arises when a HK protein undergoes a huge number of TSIs. The instance emphasizes that tissue-specific interaction purposefully drives the evolution of disordered binding regions within HK proteins.

3.5 Multi-domain and single domain housekeeping proteins

We have observed from the existing literature that the number of unique protein domain influences the pattern of protein interactions. On this ground, we have studied two aspects – i) if there exists any relation between the number of unique domains and TSI index value of the human HK proteins; and ii) if the relation is affected by the presence of disordered stretches. To test this, we have categorized the entire set of HK proteins into i) P_{TSI} (HK proteins undergoing TSIs), and ii) P_{NOTSI} (HK proteins not undergoing TSI). In contrary to our expectation, the result indicates that a higher number of P_{TSI} tends to adapt a single domain rather than multiple domains (Figure S21A). We have also noticed a similar pattern within the group P_{NOTSI}. The trend of the result remains similar even when we have categorized the HK proteins as: i) P_{HTSI}, ii) P_{LTSI} and iii) P_{NOTSI} (Figure S21B). The outcome is apparently convincible in the cases of

P_{NOTSI} and P_{LTSI}, since they do not have the urge to bind to a large number of diversified molecular targets. Therefore, they are evolutionarily designed to evolve with a fewer number of domains. On the other hand, the set of P_{TSI} or P_{HTSI} interacts with a number of structurally diversified interacting proteins. Thus, it is likely for them to adapt multiple domains. In order to address this issue, we have further questioned whether the presence of structural disorder compensates the coexistence of multiple domains within HK proteins mediating a high number of TSIs.

We have observed that SD (single domain) proteins within P_{HTSI} are significantly enriched in structural disorder (i.e., the number of disordered residues and the length of disordered regions) compared to MD (multi-domain) proteins (Figures S22). However, there is not much difference in the distribution of the number of disordered regions between SD and MD proteins in P_{HTSI} (Figure S22). On the other hand, the differences in the distributions of various parameters estimating structural disorder (i.e., number of disordered residues, length of disordered regions, and number of disordered regions) are not significant between SD and MD proteins in other groups of HK proteins (i.e., P_{LTSI} and P_{NOTSI}) (Figures S22 B and C). The result indicates that HK proteins mediating more number of TSIs, prefer adapting highly disordered single domain, rather than well-structured multiple domains. However, the result exhibits no such significant difference in the fraction of disordered regions overlapping with the protein domains between the groups of MD and SD proteins, when the entire set of HK proteins is taken into account. Hence, a single domain is not essentially needed to be structurally disordered in nature. A high degree of tissue-specific interactions rather favors the evolution of structurally disordered regions, particularly in SD proteins. It establishes that the selection pressure acts economically on the disordered stretches. Furthermore, we have noticed that three groups of P_{HTSI}, P_{LTSI}, and P_{NOTSI}

have exhibited a small but significant difference in the distribution of the fraction of disordered regions overlapping with protein domains (Figures S23). It suggests that P_{HTSI} has relatively a greater extent of the overlapping portion between the stretches of intrinsically disordered and domain regions compared to the other two groups - P_{LTSI} and P_{NOTSI}.

3.6 Housekeeping domain tends to be intrinsically disordered in nature

We have further concentrated only on the set of protein domains rather than an entire protein. Protein domains are considered to be the actual site for physical interactions, and therefore, the study of protein domain will also throw light on the type of their protein interactions. We have used RNA-Seq datasets to categorize the whole set of protein domains into "housekeeping" (HK) and "tissue-specific" (TS) following the criteria given by Lehner and Fraser (Lehner and Fraser 2004) (Section 2.6). Consequently, it has resulted in 713 HK and 919 TS domains. Moreover, we have calculated the distribution of the disordered contents (in percentage) within the group of domains, and found that the domains which are being expressed ubiquitously and resides within the human HK proteins, are intrinsically more disordered in comparison with TS domains ($P = 1.0\times10^{-32}$ according to Mann-Whitney U test). Moreover, the distribution of the sets of human HK and TS domains has shown that comparatively a higher number of HK domains (94 out of 713 domains, i.e., ~13%) are intrinsically disordered compared to TS domains (72 out of 919, i.e., ~7%), whereas the trend is different in case of well-structured domains. Well-structured domains have shown a higher enrichment within TS domains (847 out of 919 domains, i.e., ~93%) compared to HK domains (619 out of 713 domains, i.e., ~87%). The above distribution is significant ($P=1.0\times10^{-12}$), according to Fisher's Exact test. Additionally, we have found that intrinsically disordered domains (within human HK and TS proteins) are longer than well-structured domains ($P=1.0\times10^{-6}$ using Fisher's Exact test). It suggests that HK domains within

human HK proteins help in mediating TSIs, while in TS proteins, TS domains do not necessarily need to be disordered.

3.7 Literature Evidences: Role of intrinsically disordered regions within the binding sites of housekeeping proteins mediating tissue-specific interactions

In this section, we have highlighted some experimental observations supporting our hypothesis, claiming that intrinsically disordered regions are advantageous for HK proteins promoting TSIs. For this purpose, we have collected data using UniProtKB, SwissProt, Interpro databases along with experimental data taken from existing literature. For instances, RNA-binding proteins with serine-rich domain 1 (RNPS1) of human express ubiquitously (Badolato et al. 1995), and has been reported to be intrinsically disordered at their binding sites (161-240 amino acids) (Korneta and Bujnicki 2012). According to Uniport (Magrane and Consortium 2011) (http://www.uniprot.org/uniprot/Q15287) and Interpro (Mitchell et al. 2015) database (http://www.ebi.ac.uk/interpro/protein/Q15287), it has also been seen that this region holds RRM (RNA recognition motif) domain. Similarly, Polynucleotide adenylyltransferase alpha protein (PAPOA) of human has its binding site from 677 to 745 amino acids, which interacts with NUDT21. According to our dataset, nearly 55% of this region gets overlapped with the adjacent disordered stretches. Moreover, according to Uniport database, a overlapping region (508 – 743 amino acids) is seen to be serine/theorine rich. It indicates that it may be a disordered region as it has already been seen that serine rich regions are usually intrinsically disordered (Haynes et al. 2006). Buljan et al (Buljan et al. 2012) have claimed that these two human proteins have tissue-specific exons that facilitate their TSIs. Furthermore, our dataset has included the ubiquitously expressed protein Calpastatin (Calpain inhibitor), which has been reported to interact with Calpain using its flexible and disordered interacting site (135-146 amino acids). The instance has

even been supported by a few other studies (Csizmók et al. 2005; Moldoveanu et al. 2008; Fuxreiter 2012). The examples illustrate the role of intrinsically disordered regions within the binding sites or acting as a flexible linker between protein domains of HK proteins undergoing TSIs.

4. Discussion

To date, several investigations have addressed the beneficial role of unstructured proteins in mediating protein-protein interactions (Tompa and Fuxreiter 2008; Tompa et al. 2009; Schlessinger et al. 2011). In several studies, the functional and evolutionary importance of intrinsically unstructured proteins and their participation in protein-protein interactions are taken into special attention (Wright and Dyson 1999; Dunker et al. 2002; Ward et al. 2004; Tompa 2005). In this scenario, we have reported that: (i) HK proteins are significantly enriched in structural disorder compared to TS proteins in human (Section 3.1); (ii) TSIs influence the molecular evolution of structural disorder within human HK proteins and exploits their utility (Section 3.3); and (iii) utilization of structural disorder within HK proteins is energetically cost-effective for the cellular system.

While many studies have focused on the difference in several biological aspects between HK and TS proteins, we have addressed how the pattern of protein folding discriminates these two classes. We have observed that a subset of HK proteins, in contrast to TS proteins, exhibits a tendency to be in the unfolded state under normal physiological conditions (Section 3.1). According to the existing literature, intrinsically unstructured or disordered regions in a protein are shown to be highly evolvable (Brown et al. 2010; Brown et al. 2011). However, there is also a few instances where disordered stretches are highly conserved (Chen et al. 2006a; Chen et al.

2006b). We have observed that the disordered regions within HK proteins rather follow the latter instance. The disordered regions have a much lower evolutionary rate in HK proteins compared to those within TS proteins, identifying a different trend of adaptation of protein structure within the groups of HK and TS proteins (Section 3.2). Here, we have come up with a probable reason explaining our observations.

TS proteins are found to be responsible for increasing the complexity level of an organism by differentiating the types of expression and functions of different proteins in different tissues (thereby the organs) in higher eukaryotes (mammals) rather than those of the lower eukaryotes (yeast) (Stein et al. 1990a; Stein et al. 1990b). Furthermore, prior investigations have also highlighted the role of structural disorder in the emergence of organism's complexity and its capability of adapting to the environment (Schad et al. 2011; Pancsa and Tompa 2012). On this ground, we have assumed that TS proteins have gradually accumulated unstructured residues across evolutionary time, to enhance the complexity of a cellular system in a higher organism (mammals). This causes rapid evolution of disordered residues within TS proteins. In this context, in one of the recent reviews (Schlessinger et al. 2011), Schlessinger et al. have concluded their discussion with the question, "Is disorder THE major tool that simplifies the increase in complexity and adaptation to the environment?". In contrary, HK proteins are essential for the basic functions of the cellular systems and remain highly conserved across mammalian species. Nevertheless, the question of why HK proteins need to be rich in structural disorder remains an interesting but unexplored question.

To address the question, we have hypothesized that the manifestation of alteration in protein-protein interactions, within the groups of HK and TS proteins, may need to acquire specialization in the conformation of their protein structure. In this context, we have focused on HK proteins

undergoing TSIs in order to get a vivid picture of how efficiently the unstructured segments assist in protein-protein interactions (Bossi and Lehner 2009). We have observed a gradual increase in the structural disorder (i.e., number of disordered residue and length of the disordered regions) in HK proteins with the increase in TSI index. However, notably, the increase in TSI index tends to reduced the number of disordered regions (i.e., DR count) in HK proteins (Section 3.3.1).

A housekeeping protein gets expressed in all the tissues and needs to interact with different molecular targets under different tissue-specific environments. Following our definition (Section 2.2), a pair of unique interaction in a particular tissue is denoted as TSI (Figure S1). Accordingly, the TSI index of a HK protein indicates the extent of variation in their unique interacting partners. Therefore, a HK protein undergoing a large number of TSIs is actually interacting with several structurally diversified molecular targets (Figure 2). This may necessitate the adaptation of structural flexibility in HK proteins.

On this ground, we have justified that HK proteins undergoing a large number of TSIs perhaps take the advantage of disordered arms to enhance their structural flexibility (Figure 4). Intrinsically disordered regions have some advantages which assist in molecular binding of a protein to multiple different targets. IDPs can bind to the same partner in two different modes bringing about two different functions, a phenomenon commonly known as “protein moonlighting” (Jeffery 2003; Jeffery 2009; Huberts and van der Klei 2010; Copley 2012). Moreover, they also overcome steric hindrance with the aid of their flexible arms, hence forming complementary binding interfaces. This characteristic feature eventually increases the specificity of molecular binding and lowers the affinity between the pair of interacting proteins. As a result, both the processes of association and dissociation of reacting molecules during protein-protein

interactions become accessible in a cellular system. A similar trend has also been observed by Liu et al. (Liu et al. 2009) asserting that selection pressure acts on protein stability to optimize its molecular functions. However, strikingly, we have noticed that the phenomenon of TSIs actually favors the adaptation of a few long disordered regions rather than multiple short disordered regions within HK proteins. We have suggested that a single but long disordered region may provide a large extended interaction site which helps a HK protein to bind with different molecular targets resulting in a high number of TSIs (Gunasekaran et al. 2003). In contrast, multiple short disordered regions may provide an overall structural flexibility to the protein but is not be suitable for mediating multiple interactions, actively. This explanation, however, raises a crucial question – how far do the disordered regions in P_{HTSI} render the characteristics of an interacting site ?

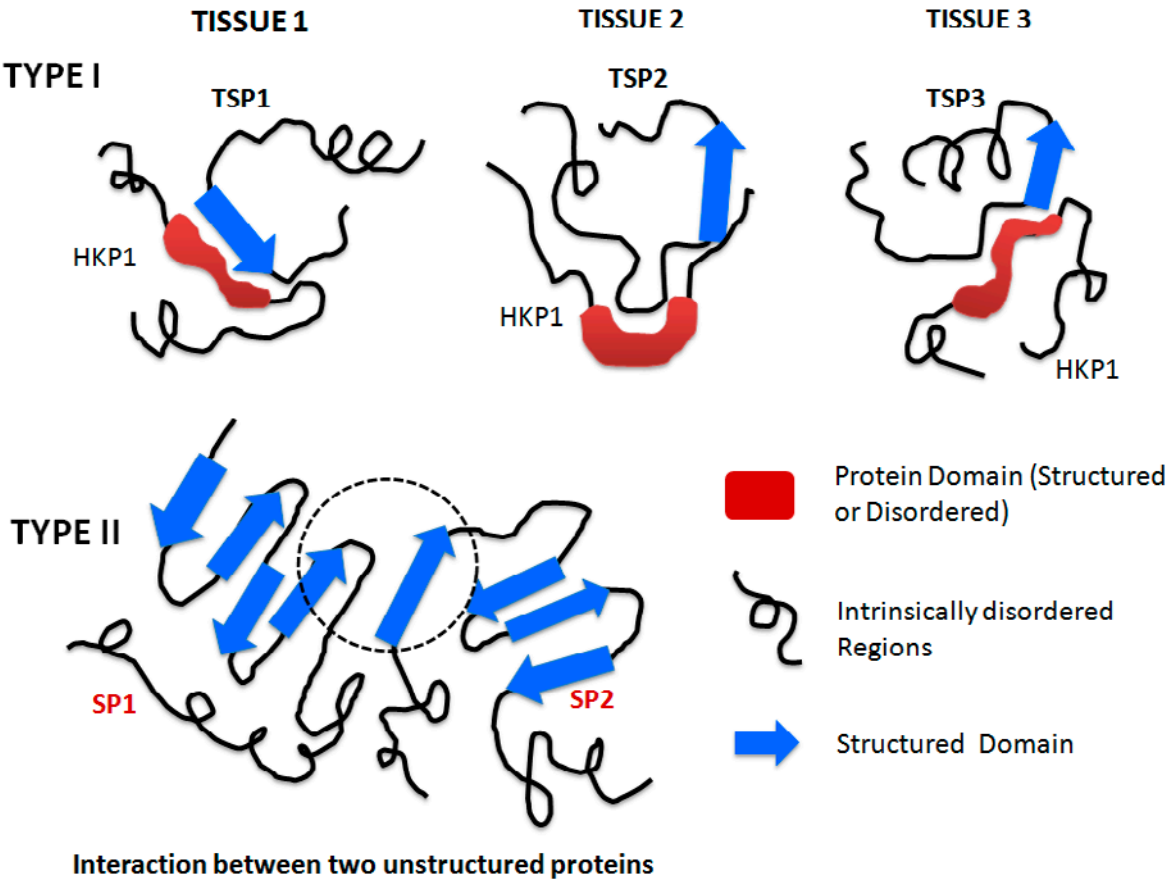


Figure 4: Role of structural disorder mediating tissue-specific interactions. TYPE I depicts three sets of molecular interactions of different TS proteins (expressed in three different tissues) with the same HK protein (HKP1). HKP1 expresses ubiquitously in all the three tissues. Here each of the interactions is considered to be TSI. HKP1 has a stretch of the intrinsically disordered segment, which can execute multiple different interactions with different structured or disordered tissue-specific proteins (TSP1, TSP2, TSP3) having mostly structured domains (shown in blue). TYPE II shows a protein-protein interaction between two highly structured proteins (SP1 and SP2) having multiple structured domains.

In the above context, we have shown that the disordered stretches in P_{HTSI} are more likely to act as MoREs and binding sites compared with other HK proteins undergoing very low or no TSI (Figure S19 and S20). Moreover, HK proteins which have a single domain, usually have a

greater extent of their domains overlapped with adjacent stretches of disordered regions (Figure S18). However, proteins with many domains do not exhibit such characteristics. Interestingly, the result has also shown that most of the HK proteins (≈ 58%) having high TSI index possess single domain (Figure S21). The observation is a bit unusual in respect to the present literature claiming that proteins, taking part in a large number of interactions, usually remain under the evolutionary pressure of adapting multiple domains or binding motifs (Kim et al. 2006b). We have thus put forward a question - Does the presence of structural disorder compensates the requirement of multiple domains in HK proteins mediating TSIs? To this end, we have reasoned that the presence of structural disorder may contribute structural flexibility to the particular protein domain, and thus help it getting into several structural conformations depending on their complimentary interacting partners (Figure 4). On the other hand, proteins having multiple domains may not need to attain a high level of flexibility. Instead, they can utilize their different domains for different interactions.

We have predicted that HK proteins, undergoing a large number of TSIs, constraint the number of protein domains and simultaneously enhance structural flexibility through acquiring disordered residues instead of maintaining multiple domains (Figure S23). It possibly balances the energy economy of the cellular system. Maintaining multiple domains, utilizing them for binding to different interacting targets, and performing different molecular interactions, may be energetically costly. Moreover, the co-occurrence of multiple domains in a single protein increases protein's molecular size, causing molecular crowding in the interacting space. On the other hand, Nussinov et al. (Gunasekaran et al. 2003) have experimented and commented that the positive selection of large extended and flexible interface, keeping the molecular size smaller, is far more economic and advantageous in the following aspects. Intrinsically disordered regions of

proteins can undergo different forms of structural conformations upon binding to its (structurally) different interacting partners. Proteins with a single domain and of flexible nature can easily maintain themselves, even in a limited interacting space. Additionally, it provides an expanded and larger surface area (for each residue), if required, for compatible molecular binding (Mészáros et al. 2007; M. Madan Babu et al. 2012).

Flexibility attained due to the lack of proper structure is believed to produce high specificity and low-affinity binding, which accelerates complex degradation enhancing the interaction speed and dynamics of signaling pathways (Dunker et al. 1998; Xue et al. 2012). These advantages allow many proteins to interact with an IDP having a single domain (Dunker et al. 1998; Gunasekaran et al. 2003; Cortese et al. 2008; Liu et al. 2009). However, a single domain HK protein, taking part in a very few TSIs or not at all undergoing any TSI, does not face a strong selection pressure in accumulating a high level of structural disorder (Figure S22 - B and C). It signifies that the adaptation of structural disorder within HK proteins is deliberately designed to promote TSIs. In contrary, TS proteins are not enriched in disordered residues. We have suggested that TS proteins have a very limited number of interactions in a few tissues, and certainly do not need to bind to a large number of structurally diverse protein targets. Therefore, the study illustrates that the evolutionary pressure acting on protein folding or unfolding is highly selective to their specific requirement in biological systems.

Conclusions

We have concluded that the necessity of mediating tissue-specific interactions has forced some HK proteins, if not all, to accumulate a high degree of structural flexibility. The presence of intrinsically disordered arm assists a particular HK protein to undergo multiple interactions in

different tissues without having multiple protein domains. Biological systems prefer maintaining single domain along with structural flexibility for HK proteins undergoing TSIs rather than the coexistence of multi-domains with a rigid structure. Besides, maintenance of multiple domains within a single protein would rather become energetically expensive. It thus sets an example of an economically balanced molecular system.

Acknowledgments

We thank Bin Xue for predicting the protein disorder using the PONDR-FIT server. We also thank Lukasz Kurgan and Chen Wang for providing the web server that predicts molecular recognition elements within disordered regions. Above all, we are thankful to the anonymous reviewers and the editor for their valuable suggestions that have helped us to improve the quality of the manuscript.

References

Babu MM, van der Lee R, de Groot NS, Gsponer J. 2011. Intrinsically disordered proteins: regulation and disease. *Curr Opin Struct Biol* **21**: 432-440.

Badolato J, Gardiner E, Morrison N, Eisman J. 1995. Identification and characterisation of a novel human RNA-binding protein. *Gene* **166**: 323-327.

Baronti L, Erales J, Habchi J, Felli IC, Pierattelli R, Longhi S. 2015. Dynamics of the intrinsically disordered C-terminal domain of the nipah virus nucleoprotein and interaction with the x domain of the phosphoprotein as unveiled by NMR spectroscopy. *Chembiochem* **16**: 268-276.

Boehr DD, Nussinov R, Wright PE. 2009. The role of dynamic conformational ensembles in biomolecular recognition. *Nat Chem Biol* **5**: 789-796.

Bossi A, Lehner B. 2009. Tissue specificity and the human protein interaction network. *Mol Syst Biol* **5**: 260.

Briggs J, Paoloni M, Chen QR, Wen X, Khan J, Khanna C. 2011. A compendium of canine normal tissue gene expression. *PLoS One* **6**: e17107.

Brown CJ, Johnson AK, Daughdrill GW. 2010. Comparing models of evolution for ordered and disordered proteins. *Mol Biol Evol* **27**: 609-621.

Brown CJ, Johnson AK, Dunker AK, Daughdrill GW. 2011. Evolution and disorder. *Curr Opin Struct Biol* **21**: 441-446.

Brown CJ, Takayama S, Campen AM, Vise P, Marshall TW, Oldfield CJ, Williams CJ, Dunker AK. 2002. Evolutionary rate heterogeneity in proteins with long disordered regions. *J Mol Evol* **55**: 104-110.

Buljan M, Chalancon G, Eustermann S, Wagner GP, Fuxreiter M, Bateman A, Babu MM. 2012. Tissue-specific splicing of disordered segments that embed binding motifs rewires protein interaction networks. *Mol Cell* **46**: 871-883.

Chang CW, Cheng WC, Chen CR, Shu WY, Tsai ML, Huang CL, Hsu IC. 2011. Identification of human housekeeping genes and tissue-selective genes by microarray meta-analysis. *PLoS One* **6**: e22859.

Chen JW, Romero P, Uversky VN, Dunker AK. 2006a. Conservation of intrinsic disorder in protein domains and families: I. A database of conserved predicted disordered regions. *J Proteome Res* **5**: 879-887.

Chen JW, Romero P, Uversky VN, Dunker AK. 2006b. Conservation of intrinsic disorder in protein domains and families: II. functions of conserved disorder. *J Proteome Res* **5**: 888-898.

Copley SD. 2012. Moonlighting is mainstream: paradigm adjustment required. *Bioessays* **34**: 578-588.

Cortese MS, Uversky VN, Dunker AK. 2008. Intrinsic disorder in scaffold proteins: getting more from less. *Prog Biophys Mol Biol* **98**: 85-106.

Csizmók V, Bokor M, Bánki P, Klement E, Medzihradszky KF, Friedrich P, Tompa K, Tompa P. 2005. Primary contact sites in intrinsically unstructured proteins: the case of calpastatin and microtubule-associated protein 2. *Biochemistry* **44**: 3955-3964.

Disfani FM, Hsu WL, Mizianty MJ, Oldfield CJ, Xue B, Dunker AK, Uversky VN, Kurgan L. 2012. MoRFpred, a computational tool for sequence-based prediction and characterization of short disorder-to-order transitioning binding regions in proteins. *Bioinformatics* **28**: i75-83.

Dosztányi Z, Csizmok V, Tompa P, Simon I. 2005. IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. *Bioinformatics* **21**: 3433-3434.

Dosztányi Z, Mészáros B, Simon I. 2009. ANCHOR: web server for predicting protein binding regions in disordered proteins. *Bioinformatics* **25**: 2745-2746.

Dunker AK, Brown CJ, Lawson JD, Iakoucheva LM, Obradović Z. 2002. Intrinsic disorder and protein function. *Biochemistry* **41**: 6573-6582.

Dunker AK, Garner E, Guilliot S, Romero P, Albrecht K, Hart J, Obradovic Z, Kissinger C, Villafranca JE. 1998. Protein disorder and the evolution of molecular recognition: theory, predictions and observations. *Pac Symp Biocomput*: 473-484.

Eisenberg E, Levanon EY. 2003. Human housekeeping genes are compact. *Trends Genet* **19**: 362-365.

Eisenberg E, Levanon EY. 2013. Human housekeeping genes, revisited. *Trends Genet* **29**: 569-574.

Farré D, Bellora N, Mularoni L, Messeguer X, Albà MM. 2007. Housekeeping genes tend to show reduced upstream sequence conservation. *Genome Biol* **8**: R140.

Fuxreiter M. 2012. Fuzziness: linking regulation to protein dynamics. *Mol Biosyst* **8**: 168-177.

Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q et al. 2011. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat Biotechnol* **29**: 644-652.

Gruet A, Dosnon M, Vassena A, Lombard V, Gerlier D, Bignon C, Longhi S. 2013. Dissecting partner recognition by an intrinsically disordered protein using descriptive random mutagenesis. *J Mol Biol* **425**: 3495-3509.
Gsponer J, Futschik ME, Teichmann SA, Babu MM. 2008. Tight regulation of unstructured proteins: from transcript synthesis to protein degradation. *Science* **322**: 1365-1368.
Gunasekaran K, Tsai CJ, Kumar S, Zanuy D, Nussinov R. 2003. Extended disordered proteins: targeting function with less scaffold. *Trends Biochem Sci* **28**: 81-85.
Habchi J, Tompa P, Longhi S, Uversky VN. 2014. Introducing protein intrinsic disorder. *Chem Rev* **114**: 6561-6588.
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**: e100.
Huberts DH, van der Klei IJ. 2010. Moonlighting proteins: an intriguing mode of multitasking. *Biochim Biophys Acta* **1803**: 520-525.
Jeffery CJ. 2003. Moonlighting proteins: old proteins learning new tricks. *Trends Genet* **19**: 415-417.
Jeffery CJ. 2009. Moonlighting proteins--an update. *Mol Biosyst* **5**: 345-350.
Kim PM, Lu LJ, Xia Y, Gerstein MB. 2006a. Relating three-dimensional structures to protein networks provides evolutionary insights. *Science* **314**: 1938-1941.
Kim PM, Sboner A, Xia Y, Gerstein M. 2008. The role of disorder in interaction networks: a structural analysis. *Mol Syst Biol* **4**: 179.
Kim WK, Henschel A, Winter C, Schroeder M. 2006b. The many faces of protein-protein interactions: A compendium of interface geometry. *PLoS Comput Biol* **2**: e124.
Korneta I, Bujnicki JM. 2012. Intrinsic disorder in the human spliceosomal proteome. *PLoS Comput Biol* **8**: e1002641.
Lehner B, Fraser AG. 2004. Protein domains enriched in mammalian tissue-specific or widely expressed genes. *Trends Genet* **20**: 468-472.
Liu J, Faeder JR, Camacho CJ. 2009. Toward a quantitative theory of intrinsically disordered proteins and their function. *Proc Natl Acad Sci U S A* **106**: 19819-19823.
M. Madan Babu, Richard W. Kriwacki, Pappu RV. 2012. Versatility from Protein Disorder. *SCIENCE* **337**.
Magrane M, Consortium U. 2011. UniProt Knowledgebase: a hub of integrated protein data. *Database (Oxford)* **2011**: bar009.
McWilliam H, Li W, Uludag M, Squizzato S, Park YM, Buso N, Cowley AP, Lopez R. 2013. Analysis Tool Web Services from the EMBL-EBI. *Nucleic Acids Res* **41**: W597-600.
Mitchell A, Chang HY, Daugherty L, Fraser M, Hunter S, Lopez R, McAnulla C, McMenamin C, Nuka G, Pesseat S et al. 2015. The InterPro protein families database: the classification resource after 15 years. *Nucleic Acids Res* **43**: D213-221.
Mohan A, Oldfield CJ, Radivojac P, Vacic V, Cortese MS, Dunker AK, Uversky VN. 2006. Analysis of molecular recognition features (MoRFs). *J Mol Biol* **362**: 1043-1059.
Moldoveanu T, Gehring K, Green DR. 2008. Concerted multi-pronged attack by calpastatin to occlude the catalytic cleft of heterodimeric calpains. *Nature* **456**: 404-408.
Mészáros B, Tompa P, Simon I, Dosztányi Z. 2007. Molecular principles of the interactions of disordered proteins. *J Mol Biol* **372**: 549-561.

Oldfield CJ, Cheng Y, Cortese MS, Romero P, Uversky VN, Dunker AK. 2005. Coupled folding and binding with alpha-helix-forming molecular recognition elements. *Biochemistry* **44**: 12454-12470.
Pancsa R, Tompa P. 2012. Structural disorder in eukaryotes. *PLoS One* **7**: e34687.
Park SG, Choi SS. 2010. Expression breadth and expression abundance behave differently in correlations with evolutionary rates. *BMC Evol Biol* **10**: 241.
Pruitt KD, Tatusova T, Maglott DR. 2005. NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res* **33**: D501-504.
Punta M, Coggill PC, Eberhardt RY, Mistry J, Tate J, Boursnell C, Pang N, Forslund K, Ceric G, Clements J et al. 2012. The Pfam protein families database. *Nucleic Acids Res* **40**: D290-301.
Receveur-Bréchot V, Bourhis JM, Uversky VN, Canard B, Longhi S. 2006. Assessing protein disorder and induced folding. *Proteins* **62**: 24-45.
Schad E, Tompa P, Hegyi H. 2011. The relationship between proteome size, structural disorder and organism complexity. *Genome Biol* **12**: R120.
Schlessinger A, Schaefer C, Vicedo E, Schmidberger M, Punta M, Rost B. 2011. Protein disorder--a breakthrough invention of evolution? *Curr Opin Struct Biol* **21**: 412-418.
Stein GS, Lian JB, Owen TA. 1990a. Bone cell differentiation: a functionally coupled relationship between expression of cell-growth- and tissue-specific genes. *Curr Opin Cell Biol* **2**: 1018-1027.
Stein GS, Lian JB, Owen TA. 1990b. Relationship of cell growth to the regulation of tissue-specific gene expression during osteoblast differentiation. *FASEB J* **4**: 3111-3123.
Tomala K, Korona R. 2013. Evaluating the fitness cost of protein expression in Saccharomyces cerevisiae. *Genome Biol Evol* **5**: 2051-2060.
Tompa P. 2005. The interplay between structure and function in intrinsically unstructured proteins. *FEBS Lett* **579**: 3346-3354.
Tompa P, Fuxreiter M. 2008. Fuzzy complexes: polymorphism and structural disorder in protein-protein interactions. *Trends Biochem Sci* **33**: 2-8.
Tompa P, Fuxreiter M, Oldfield CJ, Simon I, Dunker AK, Uversky VN. 2009. Close encounters of the third kind: disordered domains and the interactions of proteins. *Bioessays* **31**: 328-335.
Uhlén M, Fagerberg L, Hallström BM, Lindskog C, Oksvold P, Mardinoglu A, Sivertsson Å, Kampf C, Sjöstedt E, Asplund A et al. 2015. Proteomics. Tissue-based map of the human proteome. *Science* **347**: 1260419.
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.
Walsh I, Martin AJ, Di Domenico T, Tosatto SC. 2012. ESpritz: accurate and fast prediction of protein disorder. *Bioinformatics* **28**: 503-509.
Ward JJ, Sodhi JS, McGuffin LJ, Buxton BF, Jones DT. 2004. Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. *J Mol Biol* **337**: 635-645.
Warrington JA, Nair A, Mahadevappa M, Tsyganskaya M. 2000. Comparison of human adult and fetal expression and identification of 535 housekeeping/maintenance genes. *Physiol Genomics* **2**: 143-147.
Wright PE, Dyson HJ. 1999. Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm. *J Mol Biol* **293**: 321-331.

Wright PE, Dyson HJ. 2009. Linking folding and binding. *Curr Opin Struct Biol* **19**: 31-38.
Xue B, Dunbrack RL, Williams RW, Dunker AK, Uversky VN. 2010. PONDR-FIT: a meta-predictor of intrinsically disordered amino acids. *Biochim Biophys Acta* **1804**: 996-1010.
Xue B, Dunker AK, Uversky VN. 2012. The roles of intrinsic disorder in orchestrating the Wnt-pathway. *J Biomol Struct Dyn* **29**: 843-861.
Yang Z, Nielsen R. 2000. Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. *Mol Biol Evol* **17**: 32-43.
Zhang L, Li WH. 2004. Mammalian housekeeping genes evolve more slowly than tissue-specific genes. *Mol Biol Evol* **21**: 236-239.
Zhang QC, Petrey D, Deng L, Qiang L, Shi Y, Thu CA, Bisikirska B, Lefebvre C, Accili D, Hunter T et al. 2012. Structure-based prediction of protein-protein interactions on a genome-wide scale. *Nature* **490**: 556-560.
Zhao QY, Wang Y, Kong YM, Luo D, Li X, Hao P. 2011. Optimizing de novo transcriptome assembly from short-read RNA-Seq data: a comparative study. *BMC Bioinformatics* **12 Suppl 14**: S2.

Supplementary Materials

Structural Disorder: A tool for housekeeping proteins performing tissue-specific interactions

Sanghita Banerjee, Rajat K. De

Machine Intelligence Unit, Indian Statistical Institute, 203 Barrackpore Trunk Road, Kolkata 700108, India

SB: (banerjee.sanghita@gmail.com)

RKD: (rajat@isical.ac.in)

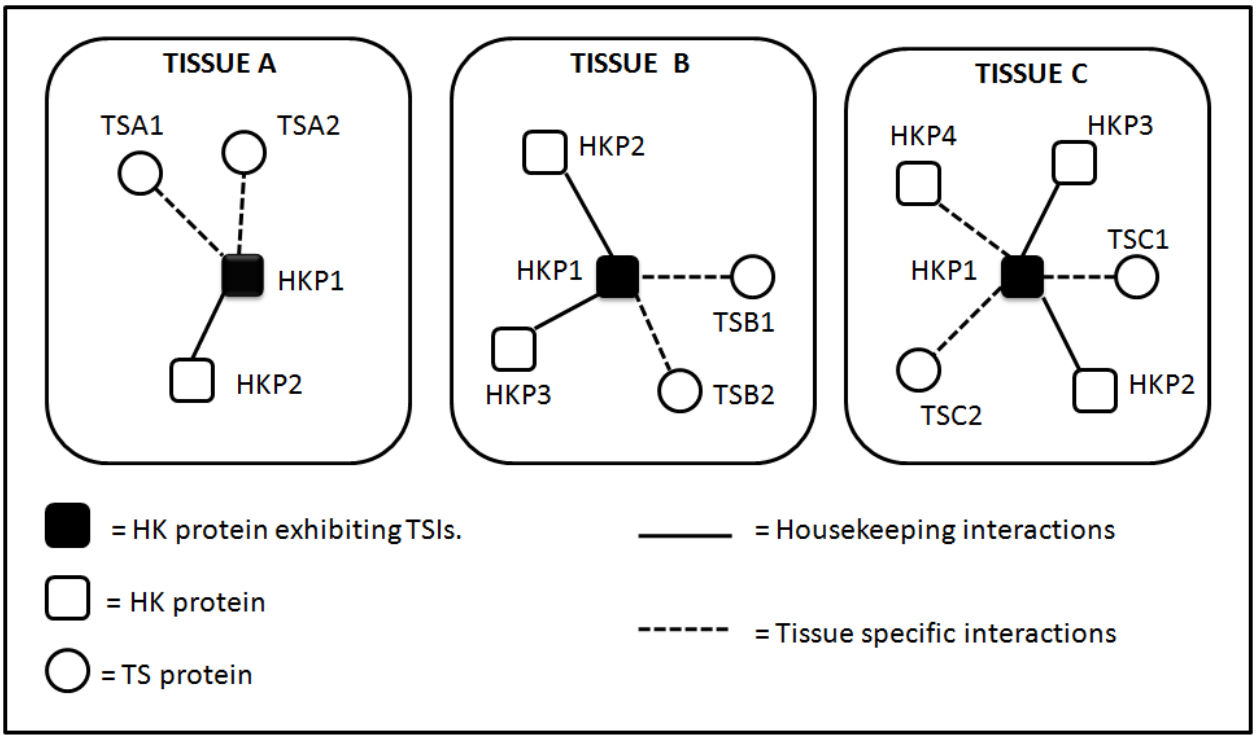


Figure S1: Tissue-specific interactions of housekeeping proteins. Tissues A, B and C represent 3 unique tissues. The schematic diagram showing the interaction network of a particular housekeeping protein (HKP1) with both of its tissue-specific interacting proteins (TSA1 and TSA2 in Tissue A, TSB1 and TSB2 in Tissue B and TSC1 and TSC2 in Tissue C) and housekeeping proteins (HKP2 in tissue A, HKP2 and HKP3 in tissue B, and HKP2, HKP3 and HKP4 in tissue C). Thus the figure depicts that the interaction between HKP1-TSA1 and HKP1-TSA2 are tissue "A" specific interactions, interactions between HKP1-TSB2 and HKP1-TSB3 are tissue "B" specific interactions, while interactions between HKP1-TSC1 and HKP1-TSC2 are tissue "C" specific interactions. In tissue C, the interaction between HKP1-HKP4 apparently seems to be tissue "C" specific according to our diagram, but being housekeeping proteins, HKP4 and HKP1 will get expressed in other several tissues, and may interact with each other in other tissues also. So the probability of the interaction, i.e., HKP1-HKP4 to be a tissue-specific interaction is very low. On the other hand, the expression of TS proteins (TSA1, TSA2, TSB2, TSB3, TSC1 and TSC2) is selective to one(or two) tissue(s), hence tissue-specific interactions between a HK and TS proteins are more meaningful.

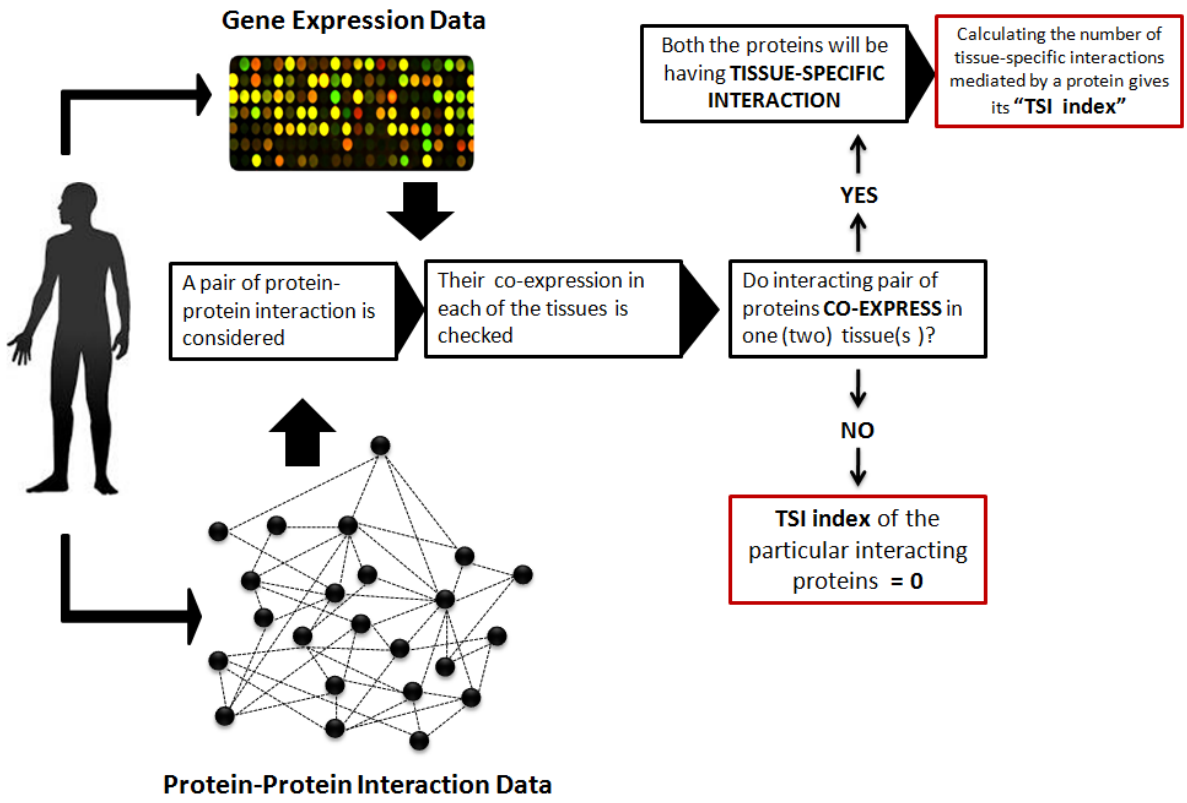


Figure S2: Steps showing the procedure of identification of tissue-specific interactions (TSIs) by integrating both gene expression and protein-protein interaction datasets of human.

EXTENDED RESULTS

Extended Results using Human Protein Atlas (HPA) dataset

1. Structural disorder in housekeeping proteins (Corresponding to Section 3.1)

Housekeeping proteins exhibit a higher enrichment in structural disorder (measured using the parameters like i. number of disordered residues, ii. number of disordered regions, and iii. length of disordered regions) compared to those of tissue-specific ones (Figure S3). It indicates that the trend of HK proteins being structurally more disordered than TS proteins is independent of the expression datasets used in the study.

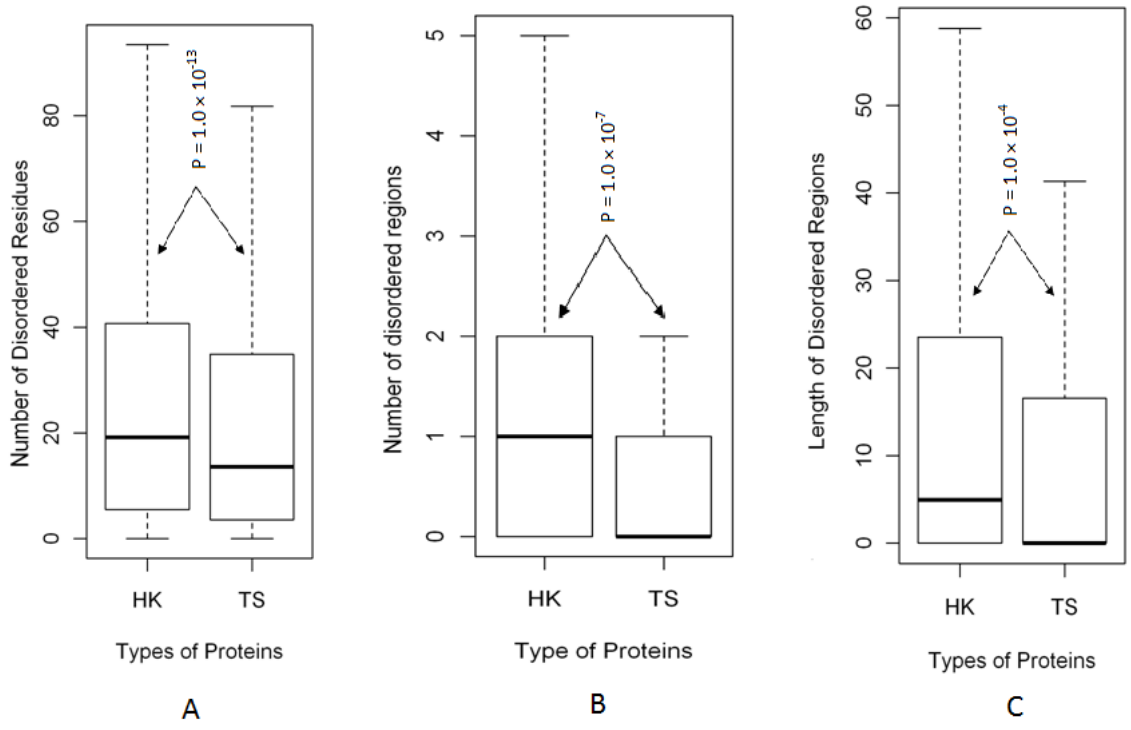


Figure S3: Difference in the A. number of disordered residues, B. number of disordered regions, and C. length of disordered regions between HK and TS proteins.

2. Evolutionary conservation of disordered regions within housekeeping and tissue specific proteins (Corresponding to Section 3.2)

The stretches of disordered regions in HK proteins evolve slowly compared to those disordered regions within TS proteins. The rate of non-synonymous (d_N) and synonymous (d_S) substitutions within disordered regions is relatively less within the stretches of disordered regions residing within HK proteins, in contrast to TS proteins (Figure S4).

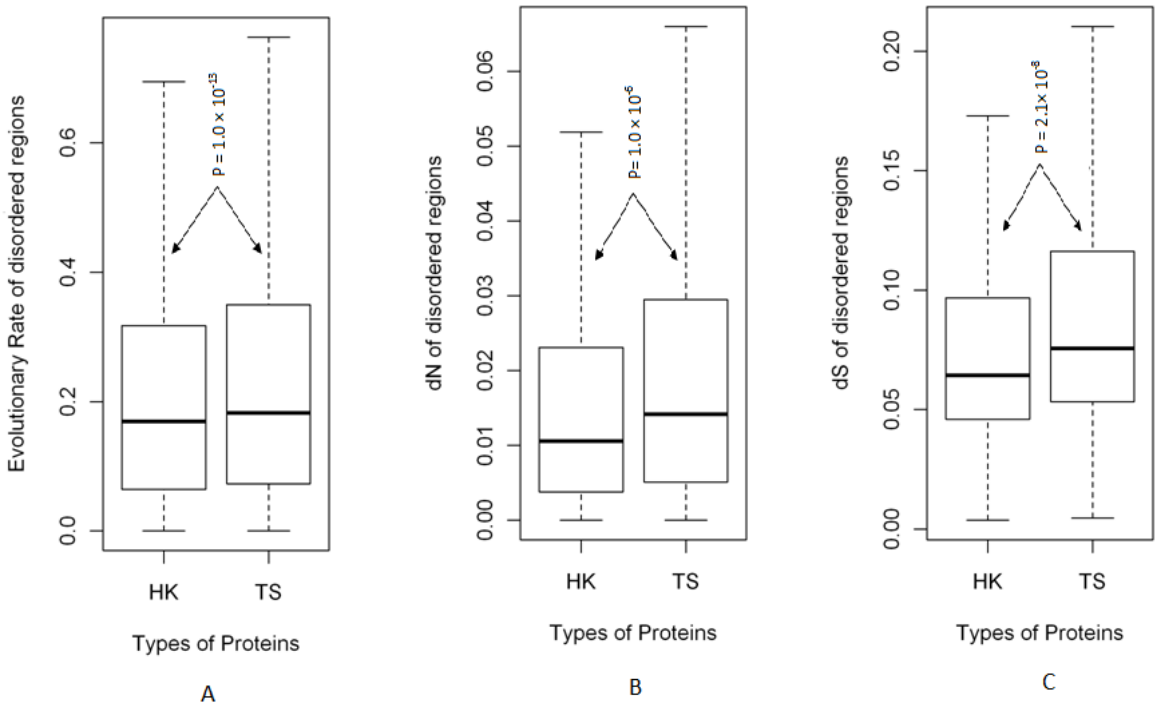


Figure S4: Differences in the distribution of A. the evolutionary rates (d_N/d_S ratio), B. the rate of non-synonymous (d_N), and C. synonymous (d_S) substitutions between housekeeping (HK) and tissue-specific (TS) proteins.

3. Influence of structural disorder in housekeeping proteins mediating tissue specific interactions (Corresponding to Section 3.3)

In the case of HPA dataset, we have categorized the HK proteins based on the threshold of average TSI index (≈ 5) as: i) P_{HTSI} (TSI index ≤ 5), ii) P_{LTSI} (TSI index > 5), and iii) P_{NOTSI} (TSI index = 0). The set of P_{HTSI} exhibits a higher enrichment in structural disorder in comparison to the other sets of HK proteins (i.e., P_{LTSI} and P_{NOTSI}).

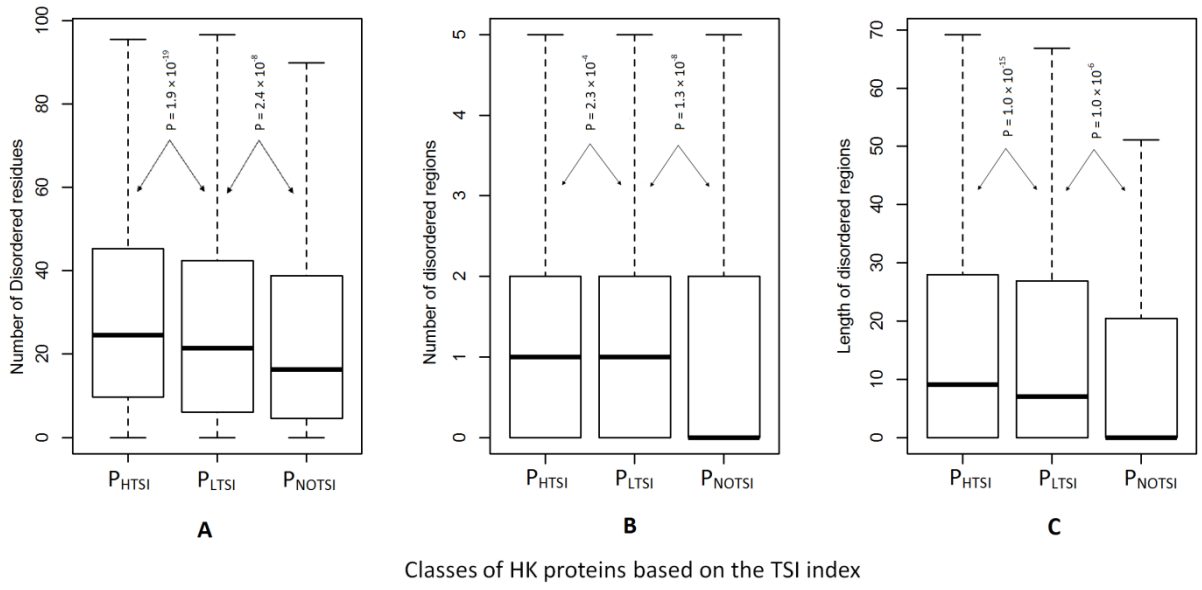


Figure S5: Differences in the distributions of the parameters – A. number of disordered residues, B. number of disordered regions, and C. length of disordered regions measured within the groups of HK proteins having varying degrees of TSI index using Human Protein Atlas (HPA) dataset.

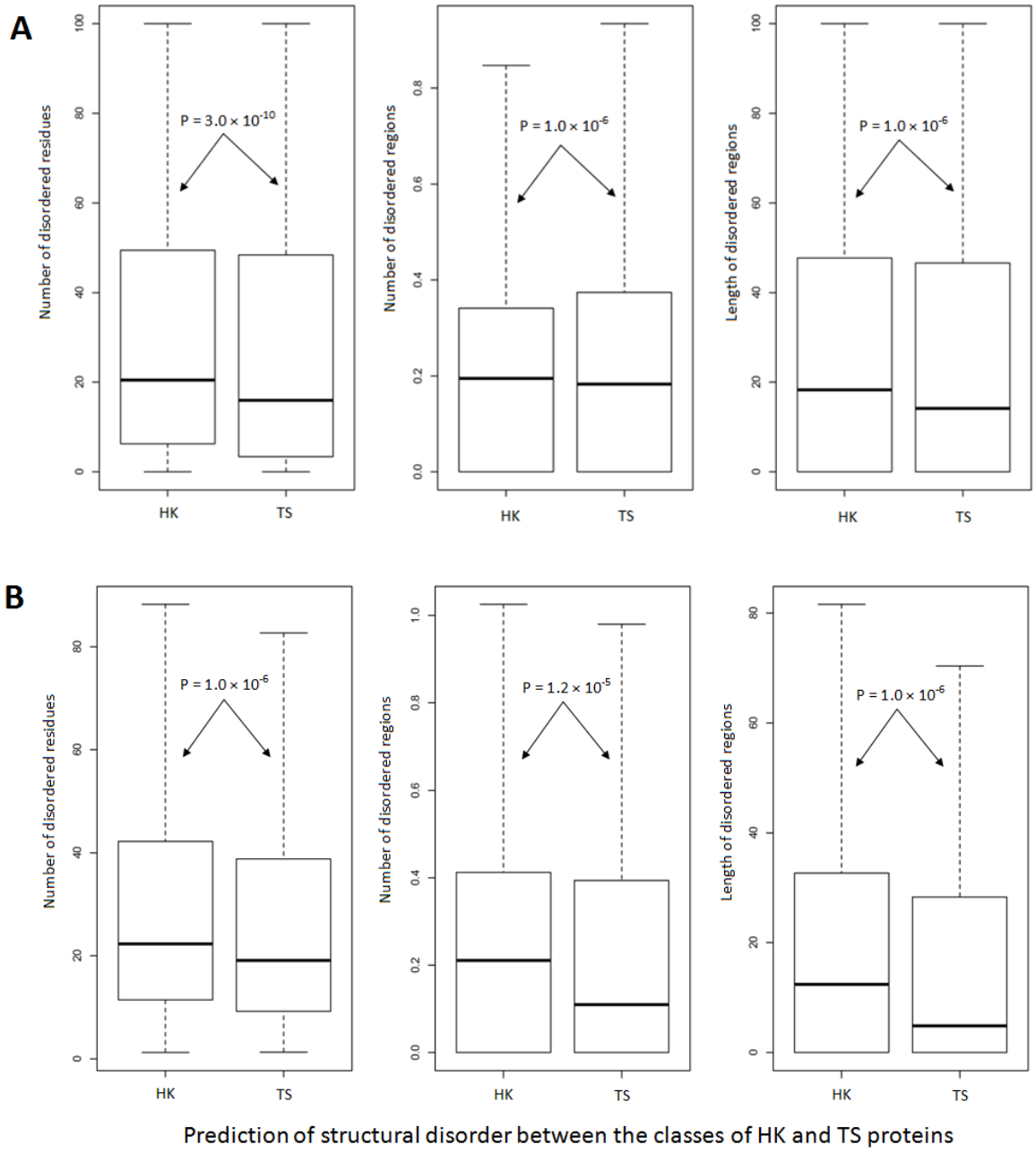


Figure S6: Prediction of structural disorder (i.e., number of disordered residues, number of disordered regions and length of the disordered regions) between the sets of HK and TS proteins using different disorder prediction tools like A) ESpritz and B) PONDR-FIT.

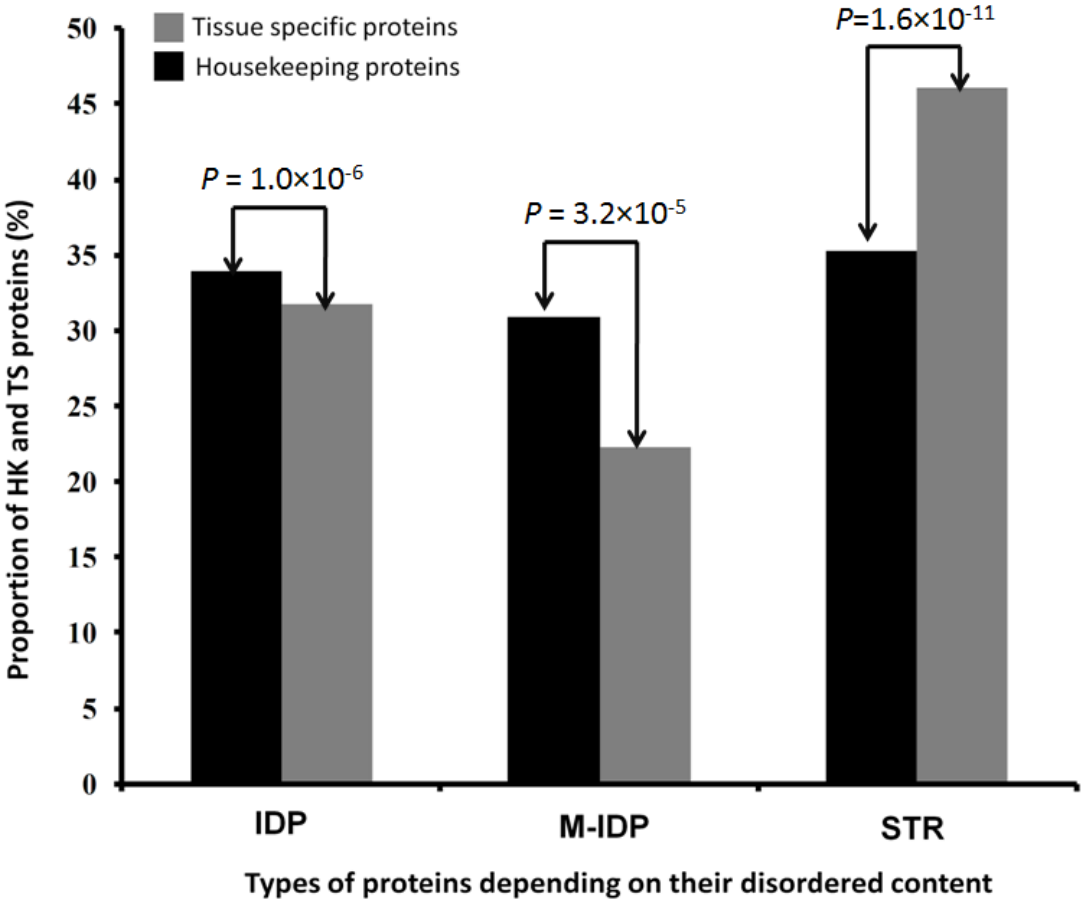


Figure S7: Bar plot showing the difference in the proportion (in percentage) of the highly disordered (IDPs), moderately disordered (M-IDPs) and well-structured proteins (STRs) within the groups of housekeeping (in black) and tissue-specific (in grey) proteins.

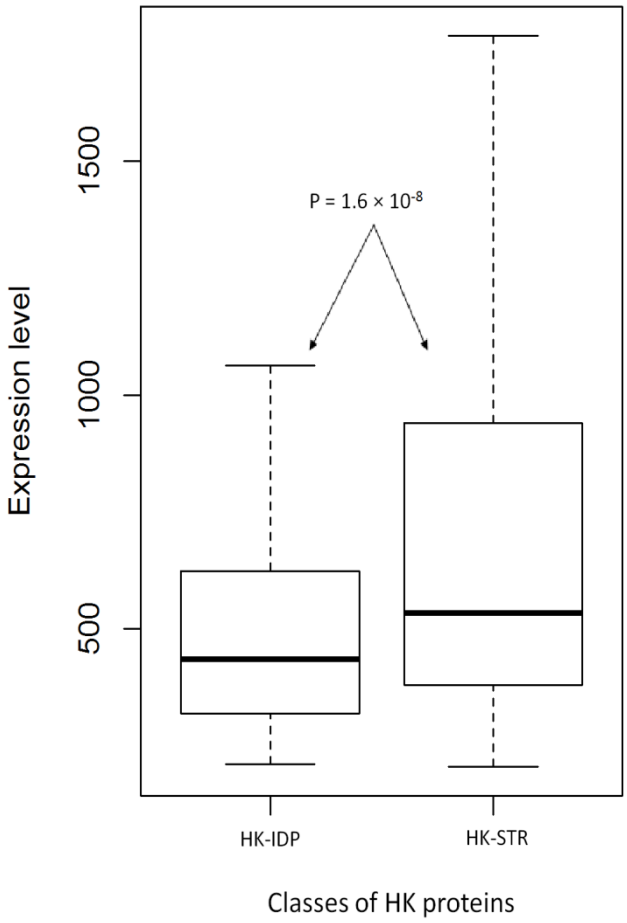


Figure S8: Boxplots showing the difference in the distribution of expression level between disordered (IDPs) and well-structured (STRs) HK proteins.

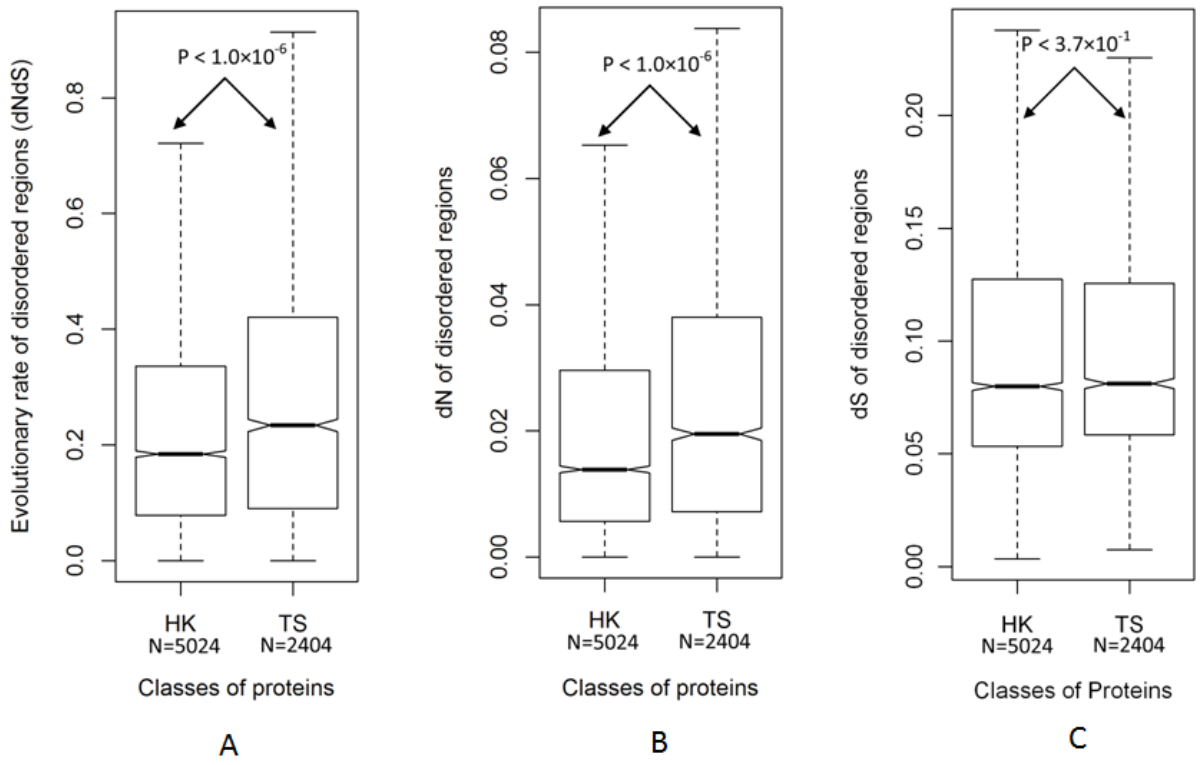


Figure S9: Boxplots showing the differences in the distributions of (A) evolutionary rate (d_N/d_S), (B) rate of non-synonymous substitution (d_N), and (C) rate of synonymous substitution (d_S) of intrinsically disordered regions between the classes of housekeeping (HK) and tissue-specific (TS) proteins.

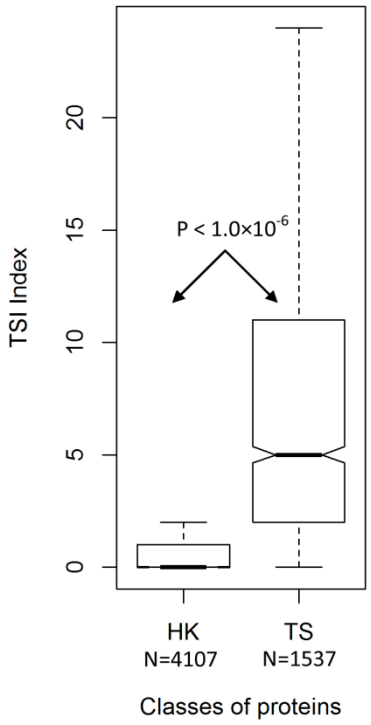


Figure S10: Boxplots showing the difference in the distribution of tissue-specific interaction index (TSI index) within the classes of housekeeping (HK) and tissue specific (TS) proteins.

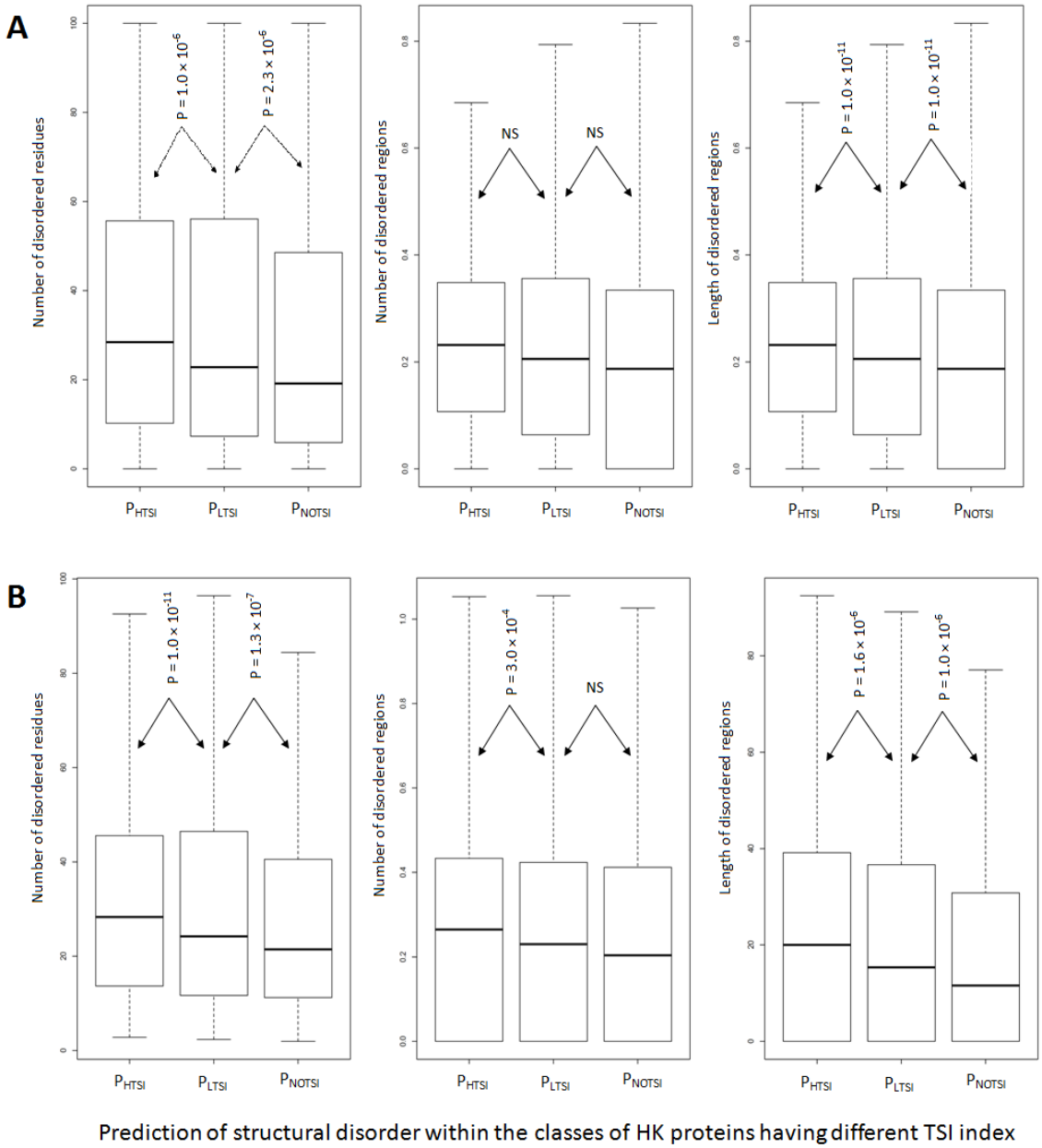


Figure S11: Prediction of structural disorder (i.e., number of disordered residues, number of disordered regions and length of the disordered regions) between the groups of HK proteins having varying degrees of TSI index (P_{HTSI}, P_{LTSI} and P_{NOTSI}) using different disorder prediction tools like A) ESpritz and B) PONDR-FIT.

Table S1: Categorization of housekeeping proteins based on flexible threshold values of TSI index.

Case No	Group	Range of TSI index	Sample Size (N)	Figure (Showing distribution of structural disorder)
I	P_{TSI}	81 to 1	1888	Figure S12
	P_{NOTSI}	0	2220	
II	$P_{HTSI\ (High)}$	81 to 11	90	Figure S13
	$P_{LTSI\ (Low)}$	10 to 1	1798	
	P_{NOTSI}	0	2220	
III	$P_{HTSI\ (High)}$	10 to 5	133	Figure S14
	$P_{LTSI\ (Low)}$	4 to 1	778	
	P_{NOTSI}	0	2220	
IV	$P_{HTSI\ (High)}$	21 to 11	47	Figure S15
	$P_{LTSI\ (Low)}$	10 to 1	911	
	P_{NOTSI}	0	2220	
V	$P_{TSI\text{-}RANDOM1}$	Grouped Randomly, Not depending on any threshold values of TSI index.	1402	Figure S16
	$P_{TSI\text{-}RANDOM2}$		1536	
	$P_{TSI\text{-}RANDOM3}$		1170	

Explanation: We have reported five cases (I, II, III, IV and V) that compare the distribution of the measures of structural disorder (i.e., number of disordered residues, number of disordered regions and length of disordered regions) among different sets of HK proteins categorized based on their TSI index values. Each of these sets is classified based on the flexible range of TSI index values mentioned in Table S1. Sample size (N), mentioned in Table S1, indicates the number of HK proteins in each of the categories.

In Cases I (Figure S12) and II (Figure S13), we have classified the entire set of HK proteins based on different TSI threshold values, and measured the differences in the distributions of various features estimating structural disorder among the groups (Figure S12). The results have shown a similar trend with the corresponding results obtained in Section 3.3. However, the

difference in the distribution of structural disorder between P_{HTSI} and P_{LTSI} in Case II is not significant due to incomparable sample size (N).

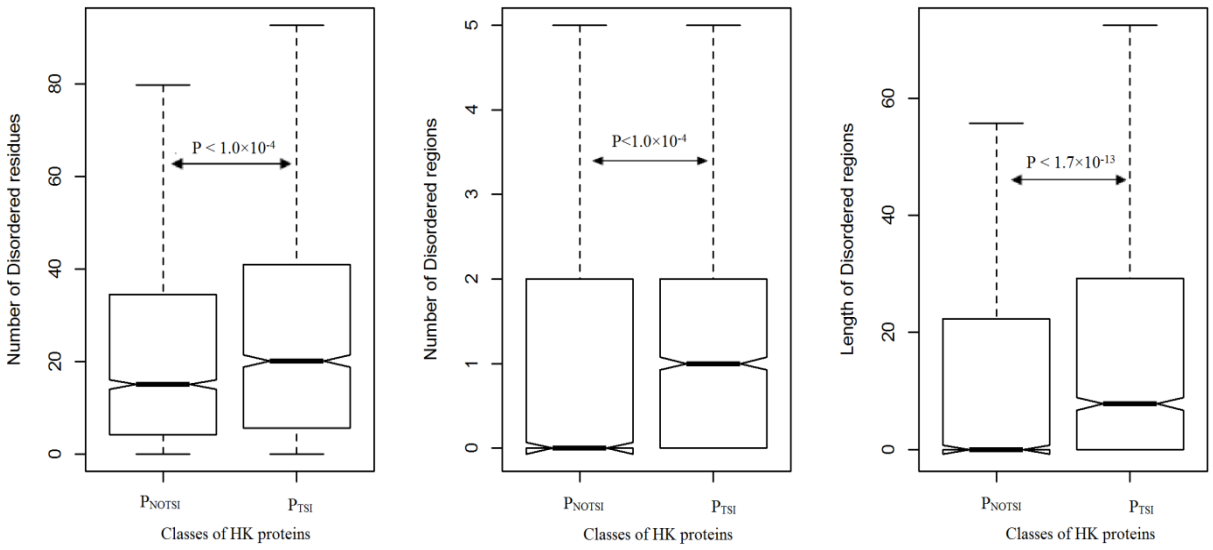


Figure S12: Boxplots showing the distribution of structural disorder between HK proteins undergoing TSIs and those not undergoing any TSIs.

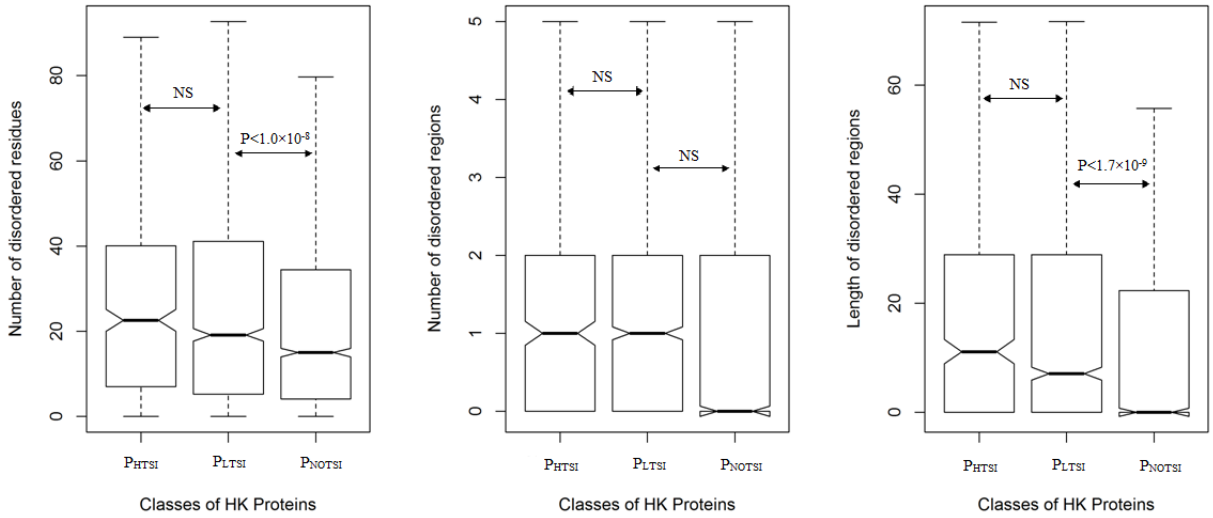


Figure S13: Boxplots showing the distribution of structural disorder between HK proteins undergoing high TSIs (P_{HTSI}), low TSIs (P_{LTSI}), and those not undergoing any TSIs (P_{NOTSI}).

In Case III (Figure S14), we have ignored the set of HK proteins having TSI index >10, as the sample size of HK proteins having TSI index > 10 is too small in comparison with the number of proteins having TSI index = 10 to 1. We have further categorized the class of HK proteins into different sets depending on the different range of TSI index and have analyzed the distribution of structural disorder among them.

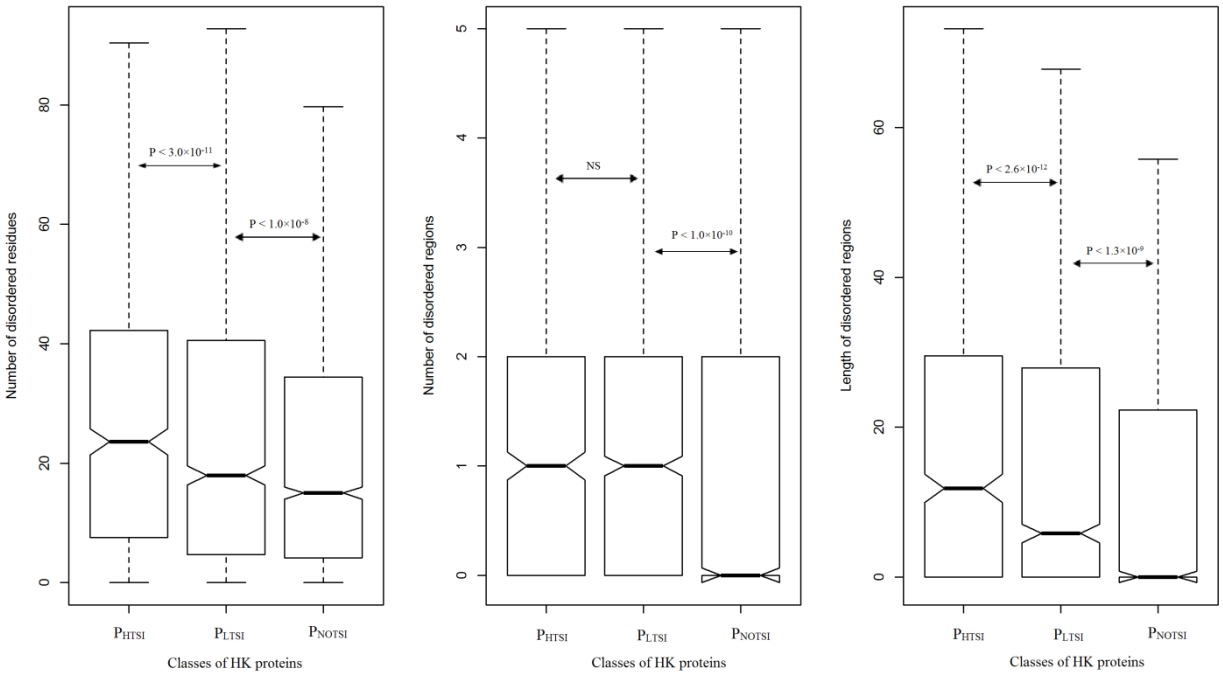


Figure S14: Boxplots showing the distribution of structural disorder between HK proteins undergoing high TSIs (P_{HTSI}), low TSIs (P_{LTSI}), and those not undergoing any TSIs (P_{NOTSI}).

In Case IV (Figure S15), we have ignored the set of HK proteins having TSI index > 21. Then, we have categorized the remaining HK proteins based on their TSI index values and compared the distributions of three parameters measuring structural disorder. The boxplots exhibit the differences in the distributions of the parameters of structural disorder (except the difference in number of disordered regions between P_{HTSI} and P_{LTSI}). However, some of the differences are not significant (NS), perhaps due to incomparable sample sizes.

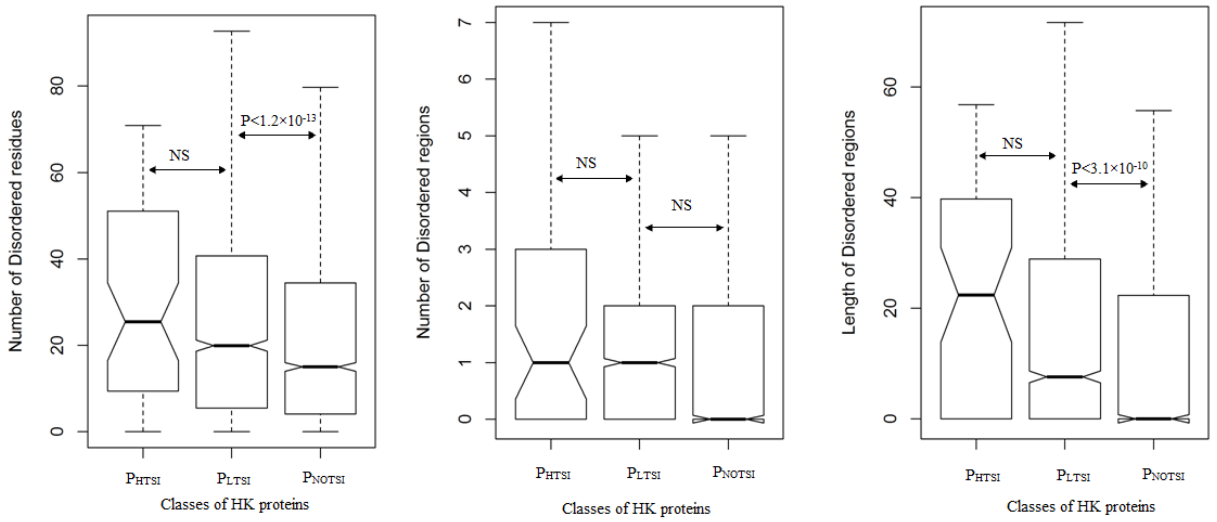


Figure S15: Boxplots showing the distribution of structural disorder between HK proteins undergoing high TSIs (P_{HTSI}), low TSIs (P_{LTSI}), and those not undergoing any TSIs (P_{NOTSI}).

In Case V (Figure S16), we have randomly grouped the set of HK proteins into three sets (TSI1, TSI2, and TSI3) of almost equal sample size. We have done this sampling in order to test whether the former grouping is at all meaningful. As expected, the differences in the distributions are not significant among the sets of TSI1, TSI2 and TSI3 (Figure S16), in spite of having comparable sample sizes. Moreover, the distribution does not even reflect any relationship between the extent of structural disorder and the TSI index of HK proteins.

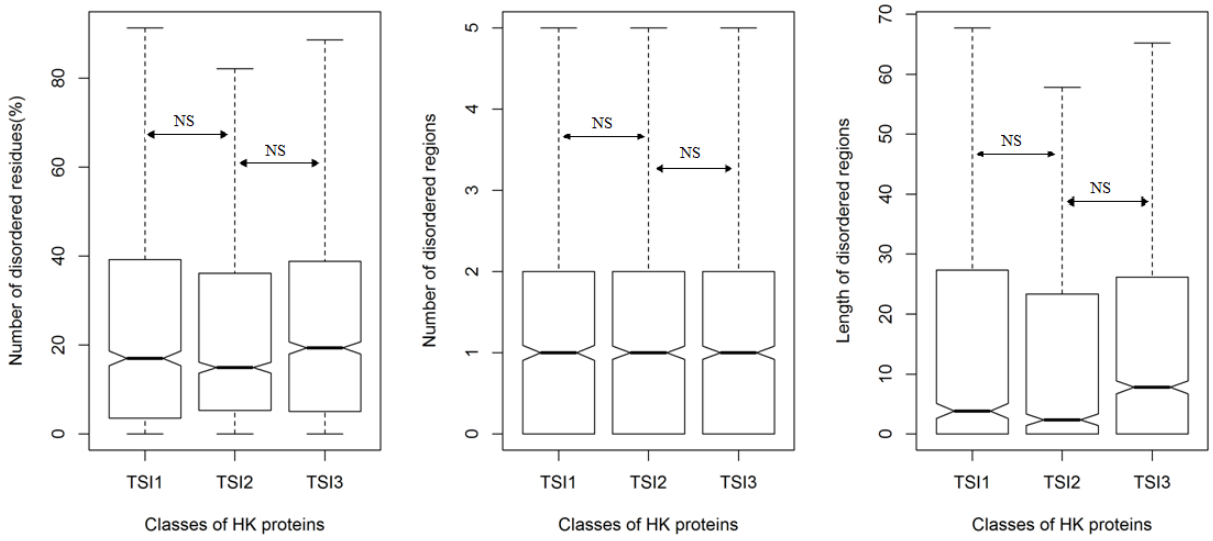
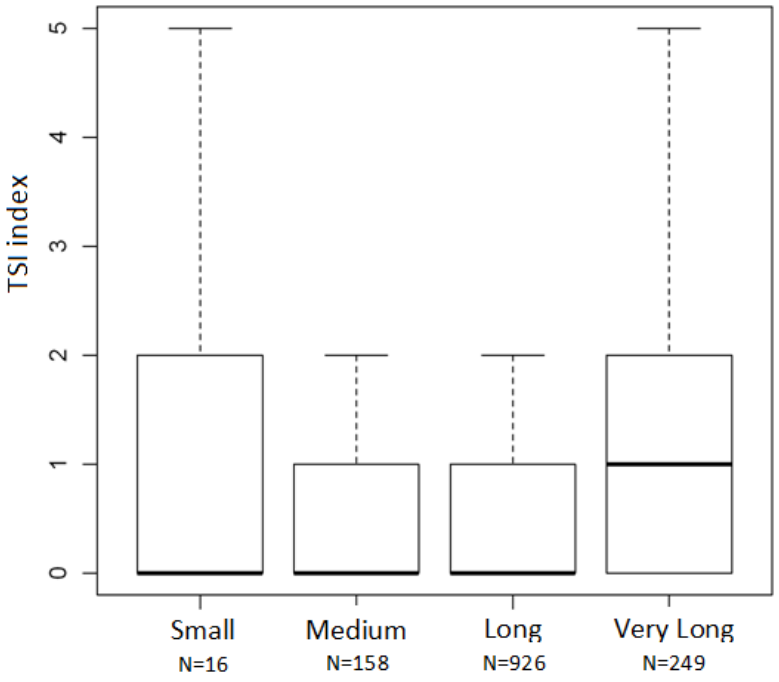
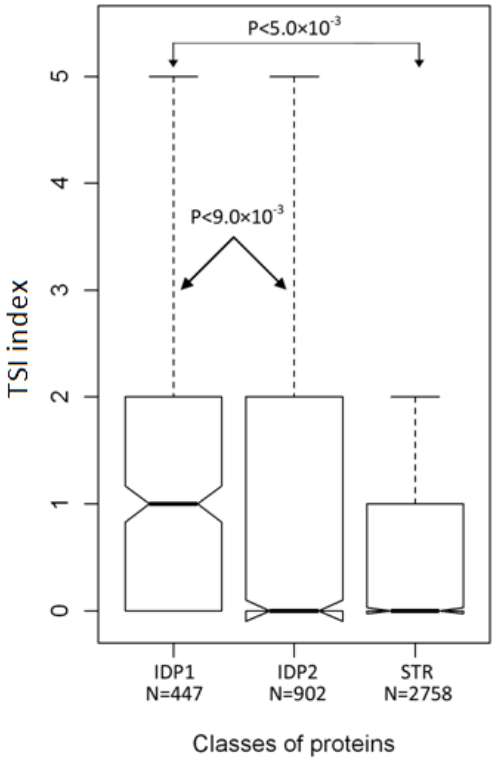


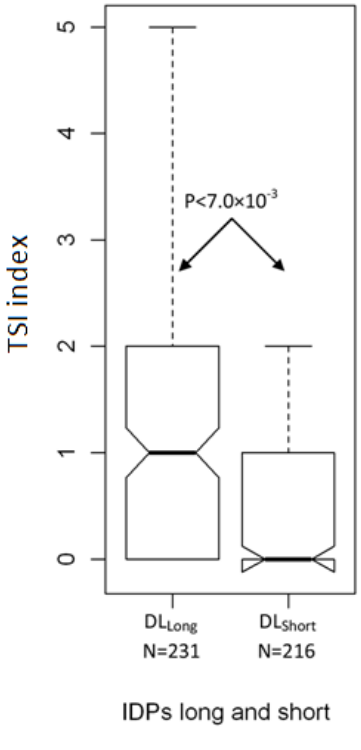
Figure S16: Boxplots showing the distribution of structural disorder between different groups of HK proteins (TSI1, TSI2 and TSI3) based on random TSI index values.



A



B



C

Figure S17: Boxplots showing A) the distribution of tissue-specific interaction index (TSI index) among the four groups (Small, Medium, Long, Very Long) of IDPs classified on the basis of the length of their disordered regions, B) Distribution of TSI index among three groups of intrinsically disordered proteins (IDPs) on the basis of the number of disordered regions (i.e., DR count) as IDP1 (DR count = 1), IDP2 (DR count > 1), and STR (DR count = 0), C) Distribution of TSI index within the two groups: DL_{Long} (DL > 120) and DL_{Short} (DL ≤ 120) categorized from the entire set of IDP1, based on the average length (≈120 residues) of the disordered region.

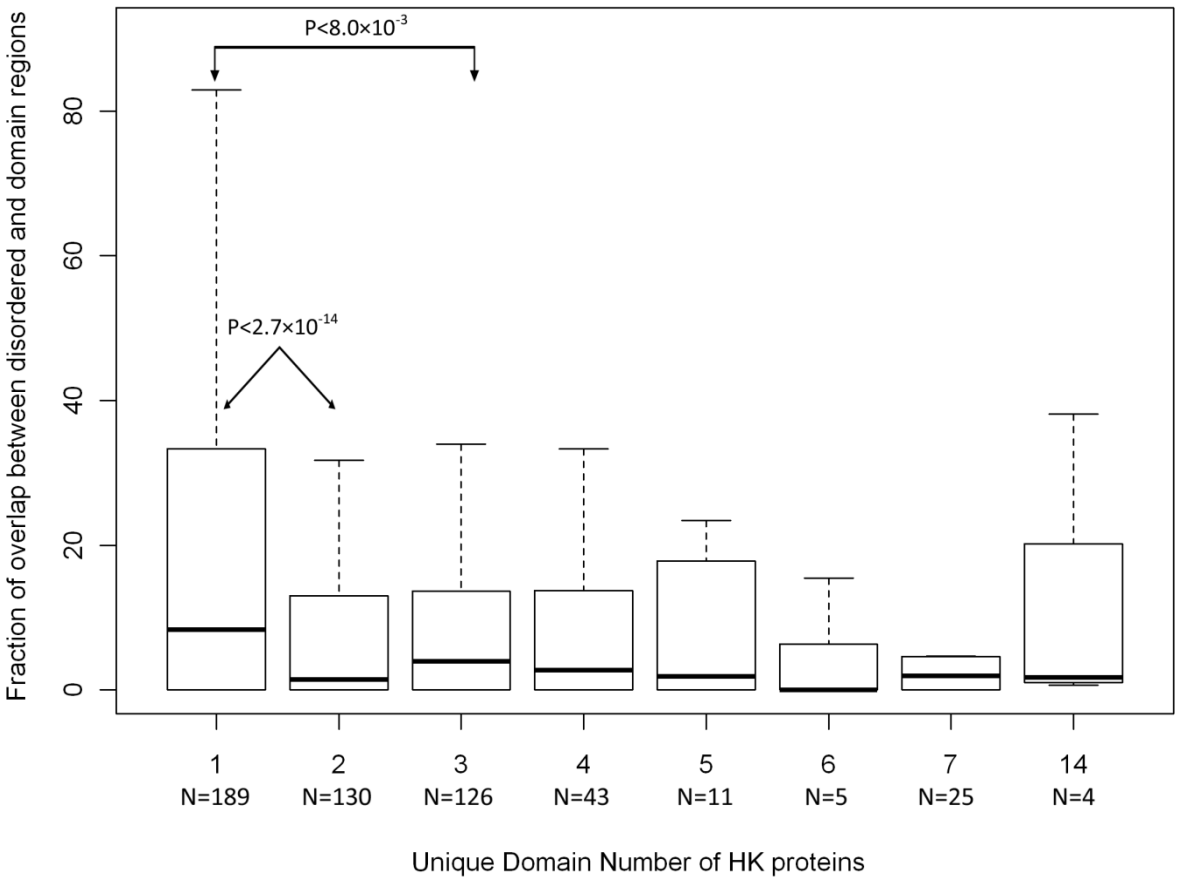


Figure S18: Boxplot showing the distribution of the fraction of disordered regions that overlaps with the adjacent protein domains among the groups of housekeeping (HK) proteins categorized on the basis of their unique domain number (ranging from 1 to 14). N denotes the number of proteins in each group.

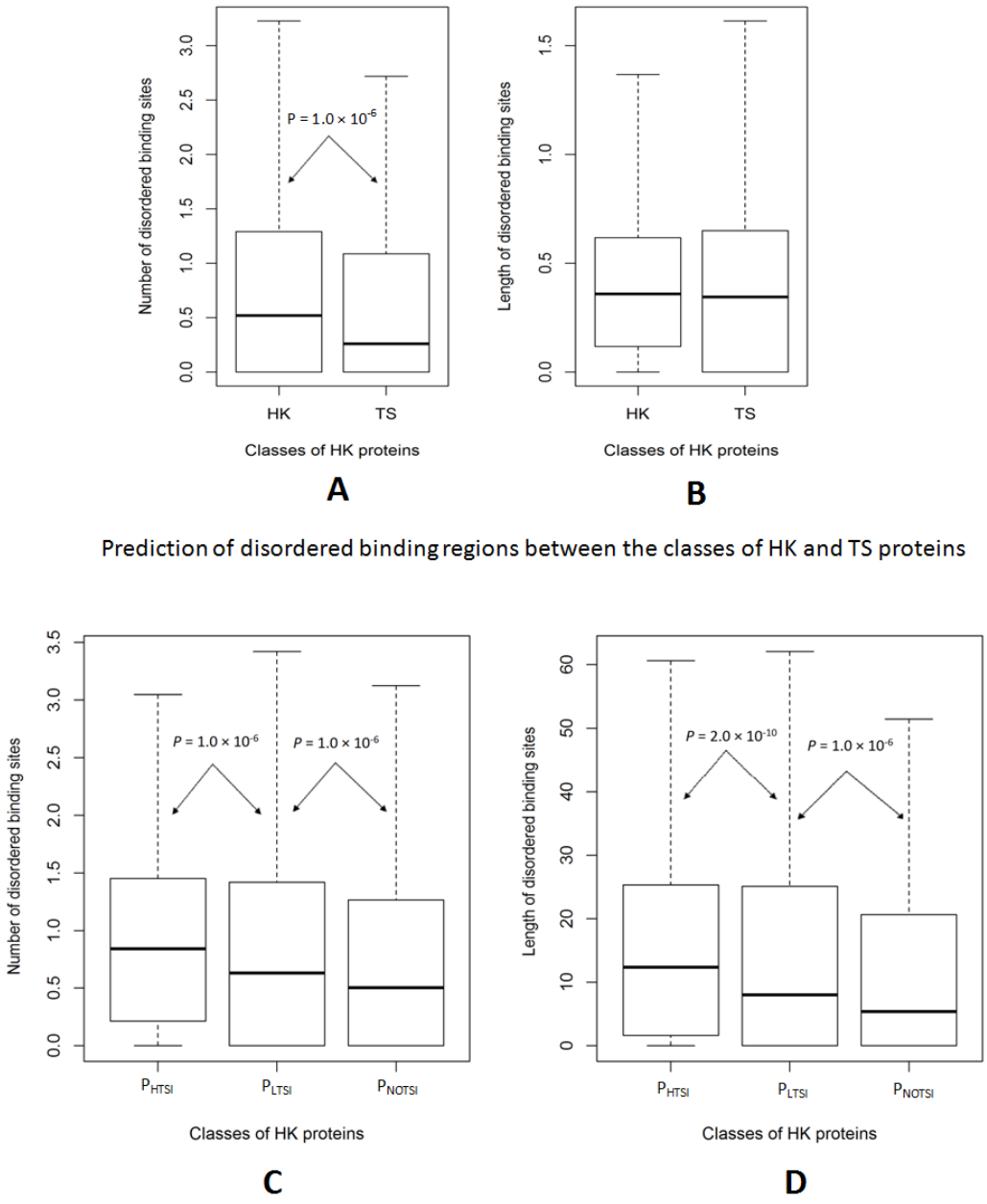


Figure S19: Prediction of disordered binding regions using ANCHOR method. Boxplots showing the differences in the distribution of the number and length of the disordered binding sites present between the classes of housekeeping (HK) and tissue specific (TS) proteins. NS stands for non-significant.

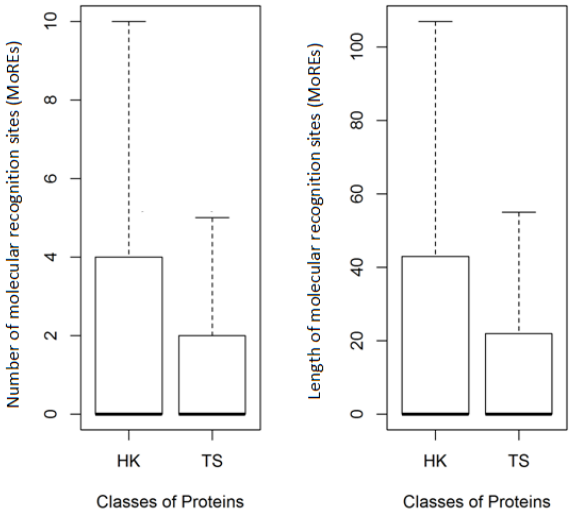


Figure S20: Prediction of molecular recognition elements (MoREs) using MoRFPred method. Boxplots showing the differences in the distribution of the number and length of the molecular recognition elements (MoREs) present between the classes of HK proteins undergoing a high number of TSIs (P_{HTSI}), low number of TSIs (P_{LTSI}) and no TSI (P_{NOTSI}).

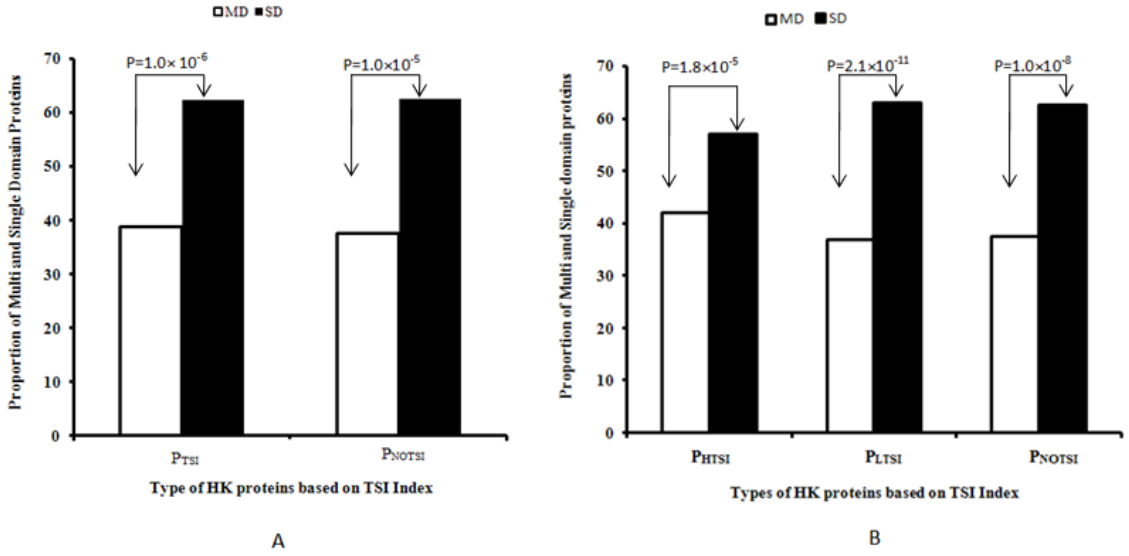
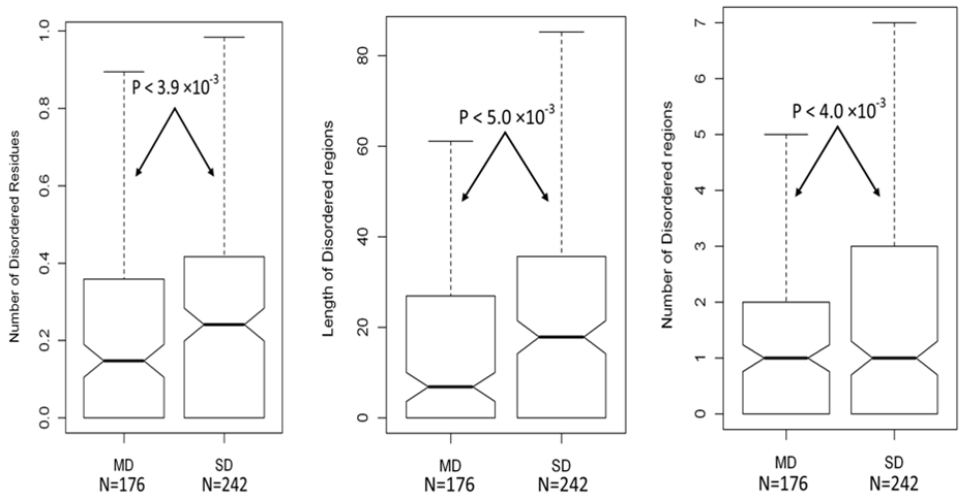
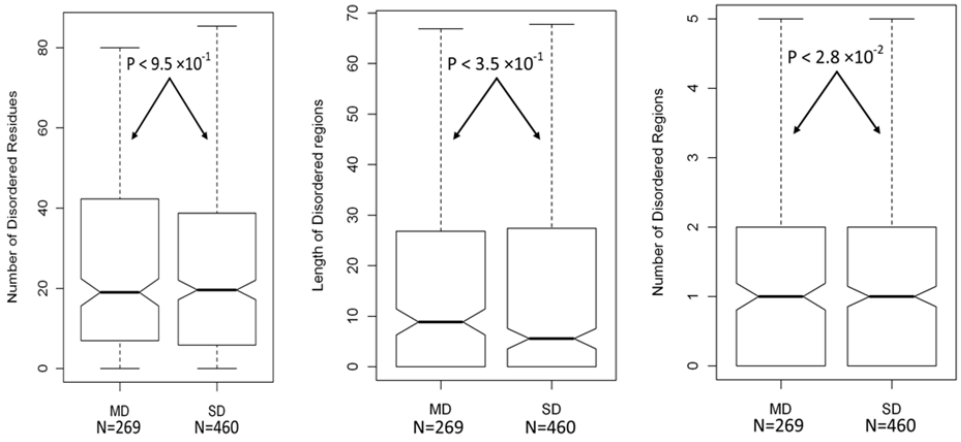


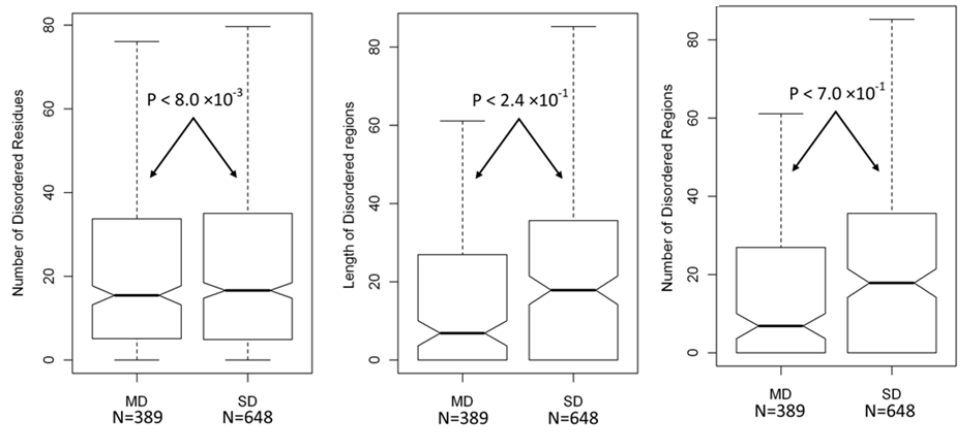
Figure S21: Histogram showing A) the proportion of multi-domain (MD) proteins and single domain (SD) proteins within the classes of HK proteins mediating TSIs and those that does not and B) the difference in the proportion of single domain and multi domain proteins within the groups of HK proteins mediating a high number of TSIs (P_{HTSI}), low number of TSIs (P_{LTSI}), and those not mediating any TSI (P_{NOTSI}).



A. Classes of proteins having High TSIs (P_{HTSI})



B. Classes of proteins having Low TSIs (P_{LTSI})



C. Classes of proteins having No TSIs (P_{NOTSI})

Figure S22: Boxplots showing the distribution of three parameters 1) number of disordered residues, 2) number of disordered regions, and 3) length of disordered regions between the groups of multi-domain (MD) proteins and single domain (SD) proteins within the classes of housekeeping (HK) proteins having A) high TSI index (P_{HTSI}), B) Low TSI index (P_{LTSI}) and C) No TSI (P_{NOTSI}).

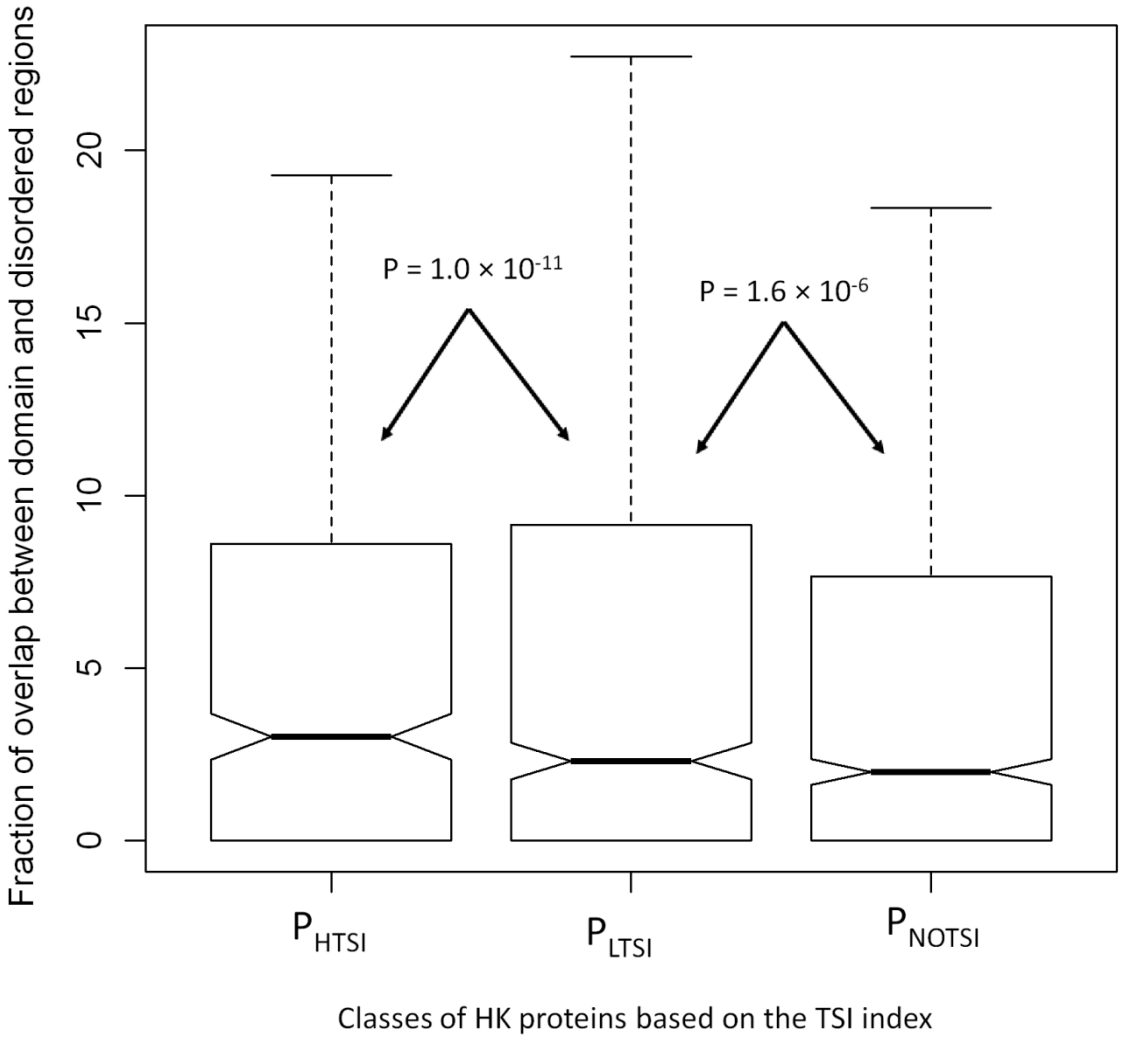


Figure S23: Boxplot showing the distribution of the fraction of disordered regions that overlap with the adjacent protein domains among the groups of housekeeping (HK) proteins having high TSI (P_{HTSI}), low TSI index (P_{LTSI}) and those not undergoing any TSI (P_{NOTSI}).