ABSTRACT  The study of unfolded protein regions has gained importance because of their prevalence and important roles in various cellular functions. These regions have characteristically high net charge and low hydrophobicity. The amino acid sequence determines the intrinsic unstructuredness of a region and, therefore, efforts are ongoing to delineate the sequence motifs, which might contribute to protein disorder. We find that PEST motifs are enriched in the characterized disordered regions as compared with globular ones. Analysis of representative PDB chains revealed very few structures containing PEST sequences and the majority of them lacked regular secondary structure. A proteome-wide study in completely sequenced eukaryotes with predicted unfolded and folded proteins shows that PEST proteins make up a large fraction of unfolded dataset as compared with the folded proteins. Our data also reveal the prevalence of PEST proteins in eukaryotic proteomes (~25%). Functional classification of the PEST-containing proteins shows an over- and under-representation in proteins involved in regulation and metabolism, respectively. Furthermore, our analysis shows that predicted PEST regions do not exhibit any preference to be localized in the C terminals of proteins, as reported earlier.

INTRODUCTION

Unfolded regions in proteins are very flexible and almost completely lack secondary structure under physiological conditions. Natively unfolded regions are implicated in multiple interactions as their structural plasticity allows them to efficiently interact with several regions. Such proteins or regions have characteristically high net charge and low content of hydrophobic amino acids. Disordered regions are enriched in P, E, K, S, and Q amino acids and depleted in W, Y, F, C, I, L, and N residues compared with an average folded protein in the PDB (Protein Data Bank). Although there are studies addressing the enrichment of different amino acids in disordered regions, only a few have reported the sequence motifs that could contribute to protein disorder.

Intrinsically unstructured regions are extremely sensitive to proteolysis as proteases cleave at sterically accessible and flexible sites. This property has been utilized to determine the location and dynamics of disordered regions in proteins. Studies on structural requirements for proteolysis also reveal the role of protein disorder and unstructured regions in calpain cleavage and ubiquitination. Recent experimental evidence shows that efficient degradation of polyubiquitinated proteins requires an additional unstructured region that serves as the initiation site for degradation and interacts with the proteasomal machinery.

The classical protein degradation targeting signals known as PEST sequences are enriched with disorder-promoting amino acids. These regions, rich in proline (P), glutamic acid (E), serine (S), and threonine (T) were first observed in rapidly degraded, eukaryotic intracellular proteins. PEST regions have a high local concentration of P, E, S, or T and to a lesser extent aspartic acid and are generally flanked by positive residues. This class of protein degradation signals confers rapid instability to many proteins. The two major protein degradation pathways implicated in PEST-mediated proteolysis are the ubiquitin-proteasome degradation and the calpain cleavage. Various experimental approaches including deletion, transfer, and mutation of PEST sequences have
shown the role and importance of PEST regions for the stability of proteins. However, deletion of PEST regions from certain proteins did not significantly affect their half-life, because PEST are conditional or regulated signals for degradation and altering the cellular environment might allow them to function as degradation signals.

PEST regions have been predicted to be solvent exposed because of their richness in hydrophilic amino acids. Rechsteiner and Rogers earlier reported conformational flexibility of PEST sequences, as these regions could not be resolved in X-ray structures. A recent report suggests that reversibly phosphorylated regions are often flexible and intrinsically disordered sections of the protein, as significant enrichment in disorder-promoting residues is seen adjoining phosphorylation sites. The consensus sites of many kinases map to PEST regions and PEST-containing proteins (PCPs) are often regulated by reversible phosphorylation. These observations suggest that PEST sequences might often be present as unstructured regions of proteins.

In the present study, we have revisited the premise that PEST regions are intrinsically disordered by analyzing the distribution of PEST sequences in experimentally characterized disordered/globular regions and the PDB chains containing PEST regions. We also evaluated the contribution of PEST sequences to overall protein disorder and their distribution in the completely sequenced eukaryotic proteomes.

**MATERIALS AND METHODS**

**Sequence Data**

The 146 disordered regions and 1,191 globular segment datasets were downloaded from http://compbio.iupui.edu/dunker/html/www/Datasets/O_PDB_S25_Aug99.htm, respectively. The following proteomes were downloaded from http://www.ebi.ac.uk/genomes: *H. sapiens*, *M. musculus*, *A. thaliana*, *D. melanogaster*, *C. elegans*, *S. cerevisiae*, *S. pombe*, and *C. parvum*. The *P. falciparum* and *S. cerevisiae* proteomes were downloaded from http://plasmodb.org and http://www.yeastgenome.org, respectively. *R. norvegicus* and *A. gambiae* genomes, all the bacterial genomes, and the *L. major* proteins were downloaded from ftp://ftp.ncbi.nih.gov/genomes. *D. discoideum* proteins were downloaded from http://dictybase.org. The list of human housekeeping genes was retrieved from http://www.compugen.co.il/supp_info/Housekeeping_genes.html. The non-housekeeping human genes list was obtained from Eli Eisenberg (personal communication). Protein sequences of these two sets of genes were extracted from Genbank. The list of bacterial adhesins and nonadhesins was obtained from Dr. Ramachandran (personal communication).

**Structure Data**

The 4,588 nonredundant PDB representative chains were downloaded from the PDB-REPRDB webserver by applying the following criteria: a) resolution <3.0 Å, b) R-factor <0.3, c) number of residues >40, d) % identity ≤30 or root-mean-square deviation >10 Å.

DSSP (definition of secondary structure prediction) program was used to calculate the accessible surface area (ASA) of each residue of the PEST sequences present in the unique representative PDB chains. A residue is defined to be a surface residue if its ASA is at least 25% of its nominal maximum area as defined by Rost and Sander.

**Gene Ontology Data**

Gene ontology association (GOA) files for *H. sapiens*, *M. musculus*, *D. melanogaster*, *C. elegans*, *E. cuniculi*, *S. pombe*, *A. thaliana*, and *S. cerevisiae* were obtained from http://www.geneontology.org and *P. falciparum* GOA file was downloaded from http://plasmodb.org. GOA files give the GO terms associated with each gene product. These files were used for counting the number of proteins with and without PEST for each of the GO terms present in the respective files. However, there are cases in which multiple entries for a single protein with identical GO ID were observed. In the present analysis, only unique entries were considered for calculating the over-/under-representation of GO terms for a given functional class.

**Prediction of PEST Motifs**

For our analysis, the EMBOSS program *epestfind* was used to find PEST regions in the proteins with the default cutoff pest score of +5.0. In this program, PEST regions are defined as hydrophilic stretches of amino acids ≥12 residues in length, which contain at least one proline (P), one aspartate or glutamate (E), and at least one serine or threonine (S, T). These regions are flanked by lysine, arginine, or histidine residues, but positively charged residues are precluded within the core sequence. The quality of a PEST sequence is determined by a scoring parameter based on the local enrichment of the critical amino acids and the hydrophobicity of the sequence motif. A score of ≥+5.0 obtained for a region with the *epestfind* is taken as a valid PEST motif. For all our analysis, we only considered valid PEST regions. In-house-developed PERL scripts were used for various analysis purposes.

**Prediction of Unfolded Regions**

Locally installed DisEMBL hot loop prediction with all default parameters (http://dis.embl.de) was used for local disorder prediction in the PEST-containing human proteins. The modified Uversky’s method developed by our group was applied on the complete eukaryotic proteomes to determine the number of natively folded/unfolded proteins. The IUPred program was also used for protein disorder prediction in the eukaryotic proteins. Proteins with ≥50% of the residues having a score of ≥0.5 were ascribed as unfolded proteins and the rest were categorized as folded proteins.

**N/C Terminal Definition for a Protein**

The initial 25% and the terminal 25% of the primary amino acid sequence of a protein were denoted as N and C terminals, respectively, in this study.
### TABLE I. Distribution of PEST Motifs in Disordered and Globular/Ondereg Regions

<table>
<thead>
<tr>
<th>Category</th>
<th>Total no. of regions</th>
<th>No. of regions containing PEST sequence(s)</th>
<th>Total no. of PEST sequences</th>
</tr>
</thead>
<tbody>
<tr>
<td>Disordered regions</td>
<td>146</td>
<td>26</td>
<td>47</td>
</tr>
<tr>
<td>Globular regions</td>
<td>1191</td>
<td>36</td>
<td>38</td>
</tr>
</tbody>
</table>

*This column denotes the number of disordered/globular regions, which contain PEST regions.

*This column denotes the total number of PEST regions present in all the disordered/globular regions.

---

**Statistics**

Chi-square test was applied for calculating the differences in the distribution of various protein classes. Chi-square probability of 0.05 with Bonferroni correction for multiple testing was used for determining gene ontology terms, over- and under-represented in PCPs. Web logo server was used to construct the Pro-X-Pro-X-Pro pattern from all predicted human PEST motifs.35

**RESULTS**

**PEST Regions and Protein Disorder**

**Distribution of PEST sequences in experimentally characterized disordered and globular segments**

PEST regions have been predicted to be unstructured as they are enriched in disorder promoting amino acid residues.2,3 To analyze whether PEST are enriched in disorder regions, we examined a dataset consisting of 146 disordered regions and 1,191 globular segments as described in the Sequence Data section of Materials and Methods. Only 3.03% of the globular protein segments contained PEST in comparison to 17.81% of the disordered regions (Table I). There was a significant over-representation of PEST in disordered regions in comparison to globular ones ($\chi^2 = 64.294, df = 1, P = 1.07E-15$).

**Analysis of PDB structures containing PEST regions**

The analysis of 4,588 unique PDB representative chains showed 238 PEST regions localized in 236 representative PDB chains, out of which 28 (11.8%) were completely unresolved, 62 (26.0%) were partially resolved, and 148 (62.2%) were completely resolved. The secondary structure analysis on 210 resolved PEST regions revealed that a significant number of residues was rather unstructured (69% irregular [U], 14% β-sheet [E], and 17% α-helix [H]) as compared with non-PEST regions (48% irregular [U], 20% β-sheet [E], and 32% α-helix [H]). IUPRED analysis on 148 completely resolved PEST regions revealed that unstructured regions had a great deal of overlap with predicted disordered regions (data not shown), consistent with the earlier report by Liu et al.36

The surface accessibility analysis for PEST regions was performed on 179 (i.e., ≥50% resolved) PEST motifs out of which 149 were found to be >50% surface exposed (Fig. 1).

**Occurrence of PEST sequences in predicted disordered regions**

DisEMBL was used to predict hot loops in the 9,827 human PCPs. The overlap of PEST residues (19,073 PEST regions) with local regions of disorder (hot loops) was calculated. For PEST sequences, 22.84% completely mapped within and 50.19% partially overlapped with the disordered regions.

We further analyzed the distribution of PEST motifs in eukaryotic predicted unfolded and folded proteins using the modified Uversky method32 and IUPred program33 (Fig. 2, Supplementary Table II). The results of both methods independently demonstrate that a major fraction of predicted unfolded proteins is enriched in PEST regions. In all the proteomes analyzed, a highly significant difference in the distribution of PEST proteins was observed between predicted unfolded and folded sets (Supplementary Table III).

**Abundance of PCPs in Eukaryotes**

The analysis of 11 completely sequenced eukaryotes also shows that PCPs make up a substantial fraction of the proteome (~25%) (Fig. 3, Supplementary Table IV) in comparison to prokaryotes (~5%, data not shown). P. falciparum and E. cuniculi were exceptions as they have low PCP representation (~11%). This might not be related to their intracellular parasite status, as proteins of partially sequenced L. major and C. parvum showed a normal trend (28% and 21%, respectively) of PCP distribution. The extremely high AT% of P. falciparum genome is also not related to this observation, as other AT-rich organisms such as D. discoideum (24%) and C. parvum (21%) also followed the general trend.

**Functional Classification of PCPs**

Classification of eukaryotic PCPs into Gene Ontology classes indicates their over-representation in categories related to transcription regulation, signaling, endocytosis, and under-representation in metabolism and transport.
proteins (Supplementary Table V). PCPs were also over-represented in human non-housekeeping as compared with housekeeping proteins [37% compared with 28% \( \chi^2 = 16.301, df = 1, P = 0.0001 \)].

The classification of \( P. falciparum \) PCPs revealed that about one-third of all PEST regions was present in 54 PfEMP-1 family of extracellular proteins (Supplementary Table V). “Cysteine type proteases” was the only intracellular class that showed significant over-representation of PEST regions.

Positional Preference of PEST Motifs in N/C Terminal of a Protein

PEST sequences have been reported to be preferentially present in the C terminal region of the proteins.\(^{16}\) Although the evidence for N and C terminal PEST regions in proteolysis has been experimentally characterized,\(^{37,38}\) we observe that reports of C terminal PESTs have been more prevalent. To analyze for any preferential localization of PEST sequences within a protein, we mapped 19,071 PEST sequences in 9,827 human PCPs. There were 4,396 PEST regions mapped to the N terminals as compared with 4,638 in the C terminals. The data show that there is no preferential localization of PEST regions in the C terminus of the protein. We also analyzed for differences in scores of PEST motifs belonging to N and C terminal regions. \( P \) value of the two-tailed, unpaired \( t \) test for this analysis was found to be insignificant (\( P = 0.1 \)).

DISCUSSION

PEST regions have been extensively studied as protein degradation signals and their role as phosphorylation targets and protein–protein interaction sites is also reported.\(^{16,19,24,25,39}\) These regions are rich in disorder-promoting amino acids and, hence, have been predicted to be intrinsically unstructured. In this study, we have analyzed both the role of PEST regions in protein disorder as well as their distribution in the eukaryotic proteomes. This analysis carries implications for the various functions attributed to the PEST regions.

Role of PEST Regions in Protein Disorder

The abundance of intrinsically disordered proteins and segments in eukaryotes as compared with prokaryotes is well reported.\(^{40}\) Studies enumerating the contribution of sequence motifs to the increased disorderedness in eukaryotes are limited.\(^{7}\) Our data reveal that PEST sequences are over-represented in characterized, nonredundant disordered regions (Table I). The low occurrence of PEST motifs in representative PDB chains indicates that these regions might be predominantly disordered (Supplementary Table I). The prevalence of irregular secondary structure and surface-accessible residues in the PEST regions in PDB further indicates their intrinsic unstructuredness. Predicted unfolded proteins in eukaryotes also show a significant enrichment of PEST regions as compared with the folded ones (Fig. 2, Supplementary Table I).
II). The prevalence of PEST regions in eukaryotes compared with prokaryotes might partially explain the presence of a large number of disordered segments in the former class.

A recent study reports the enrichment of a proline-rich motif and a charged pattern (positive or negative) in disordered regions. In the identical datasets studied by Jones et al., we observed an enrichment of PEST sequences in disordered regions as compared with the globular segments (Table I). Furthermore, we examined the Pro-X-Pro-X-Pro motif composition in all the predicted human PEST regions using WebLogo server. The composition of proline-rich motif Pro-X-Pro-X-Pro in PEST sequences was similar to the pattern seen in disordered regions (Supplementary Fig. 1). The frequency of the negative charged patterns in the PEST regions is also shown in Supplementary Table VI. Thus, PEST regions seem to be one of the longer motifs, which might have a role in intrinsic unstructuredness and identifying these regions in proteins may well be advantageous for local disorder prediction.

Unfoldedness and surface accessibility might augment the function of PEST regions as degradation signals, phosphorylation targets, etc. Unfolded regions are shown to be extremely labile to proteolytic cleavage and proteins with unstructured regions have a shorter half-life. This observation might be explained by the abundance of PEST regions in the disordered sequences.

Significance of PEST Regions and Protein Disorder in Proteolysis

The function of PEST regions as protein degradation signals is well known and cleavage by calpain proteases and ubiquitin-proteasome machinery in these sites is extensively reported. A recent report shows caspase-mediated cleavage in a PEST motif of human histone deacetylase4 protein. Furthermore, as PEST sequences are conditional degradation signals, their function can be regulated by phosphorylation and protein binding.

In the ubiquitin-proteasome pathway, a protein is marked for degradation by addition of multiple ubiquitin molecules to the targeted protein. The first ubiquitin molecule is added to a specific internal lysine residue of the protein and the additional ubiquitin molecules are added to conserved lysine residues within ubiquitin. The lysine residues bordering the PEST regions can form the substrates for ubiquitination machinery. The polyubiquitinated protein is then transported to the proteasome complex where it is cleaved into oligopeptides.

Calpains are calcium-dependent proteases and Rogers et al. hypothesized that the negatively charged residues in the PEST regions could sequester calcium for activating calpains. A study on 106 calpain cleavage sites of 49 substrates shows that the preferred residues at P1 (cleavage site) are Lys, Tyr, and Arg and P2 are Leu, Thr, and Val. It was also shown that the 11 residues around the scissile bond showed sequence preferences among which several of them were for P, S, and T. Because PEST regions are flanked by positive charges (Lys, Arg) and enriched in P, S, T residues, these regions can be targeted for calpain cleavage. Tompa et al. also reported that the position of the PEST region in calpain substrates was nonrandom with respect to the calpain cleavage site as most often it occurred at the C terminal end of the cleavage site.

Caspases, another family of proteases, cleave their substrates on the carboxyl side of an aspartate residue. The presence of aspartic acid residues in the PEST region can make a potential caspase cleavage site as seen in human histone deacetylase4. A recent in silico analysis of 280 caspase substrates shows that 55.6% of these sites are localized within PEST regions. Thus, PEST regions function as targets for multiple protein degradation systems.

Proteases generally cleave their substrates in sterically accessible and flexible regions. The extreme sensitivity of natively disordered regions to proteolytic cleavage might be attributed to their surface accessibility and flexibility. Several studies on the structural requirements of proteolysis also show that proteolytic cleavage predominantly occurs at unstructured regions. Secondary structure analysis of the 106 calpain substrates showed that the enzyme predominantly cleaves in the disordered regions. The C terminal flanking decapetide region from the calpain cleavage site, wherein PEST regions are preferentially found, was also shown to be highly unstructured (~90%). Analysis of 135 ubiquitination sites in 95 proteins of S. cerevisiae showed that almost all the ubiquitin-modified lysines were surface exposed and accessible and the ubiquitination sites showed a statistically significant preference within loops. Recently, a report on secondary structure prediction in the 280 caspase substrates (~55% PEST regions) also showed that predominantly these sites were located within random coils. Our analysis on PDB chains containing PEST regions showed that a major fraction of PEST residues were unstructured and surface accessible. These evidences indicate the importance of protein disorder in PEST-mediated proteolysis.

Hence, the flexibility seen in PEST regions because of their intrinsic unstructuredness along with the presence of sequence features recognized by various degradation pathways make them ideal substrates for proteolysis.

Functions of PCPs in Eukaryotic Proteomes

Our proteome-wide data show an abundance of PCPs in eukaryotes, with the exception of P. falciparum and E. cuniculi. Their abundance indicates that PEST-mediated proteolysis might be more prevalent and important in eukaryotes. Over-representation of PCPs in certain functional groups such as nucleic acid and protein binding, transcriptional regulation, and signal transduction reflects the increased requirement of regulated proteolysis in these classes. Earlier studies have also reported the enrichment of PEST motifs in various regulatory classes of proteins including phosphatases, annexins, cyclic nucleotide signaling pathway proteins, EF-hand calcium-binding proteins, calmodulin binding proteins, etc.

The low incidence of PCPs in P. falciparum and E. cuniculi may indicate less dependence of these organisms...
for PEST-mediated proteolysis or involvement of an altered PEST sequence. Analysis of \( P. falciparum \) proteins revealed that PEST regions are enriched in pPEM-1 class of extracellular proteins. These proteins are transported to the membrane of the infected red blood cells and mediate their adhesion and clumping to vascular endothelium cells.\(^52\) The cell adhesion caused by these proteins leads to most malaria-related deaths. An earlier study on chromosome 2 proteins of \( P. falciparum \) also found abundance of PEST regions in this class of proteins.\(^53\) On the contrary, Rifins—another class of host cell membrane targeted proteins with unclear function—were devoid of PEST regions. Interestingly, \( H. sapiens \) also showed over-representation of PCPs in cell adhesion proteins (Supplementary Table V). We also found over-representation of PEST sequences in bacterial cell adhesion proteins as compared with nonadhesins (18% as compared with 11%, \( \chi^2 = 12.611, df = 1, P = 0.0004 \)). The data indicate a direct/indirect role of PEST-like sequences in cell adhesion processes. We found a report wherein the PEST regions in eukaryotic cadherin family of cell adhesion proteins are implicated in the binding to \( \beta \)-catenin protein that prevents proteolysis and thus has an indirect role in cell adhesion.\(^54\)

“Cysteine-type proteases” was the only intracellular class in \( P. falciparum \) with over-representation of PCPs with particularly six of the eight SERA (serine repeat antigen) proteins containing PEST regions. SERA proteins have been implicated in rupture and invasion process and have been shown to require proteolytic processing directed by trans-acting proteases.\(^55\) Presence of PEST regions may have a role in the proteolytic processing and regulation of intracellular levels of SERA proteins.

In a previous study, a small dataset of eight proteins containing functional PEST elements was analyzed and a statistically significant presence of PEST regions in carboxy-terminal extensions was shown by Rechsteiner and Rogers.\(^16\) Our proteome-wide data do not reflect any bias toward the presence of predicted PEST regions in the carboxy terminals of proteins. Furthermore, experimental data assessing the functional potential of PEST might shed light on the differences, if any, between the N and C terminal PEST sequences.

**CONCLUSIONS**

In summary, our data show that PEST regions might be one of the common sequence motifs contributing to eukaryotic protein disorder. Intrinsic unstructuredness of PEST regions in eukaryotic genomes may explain their involvement in proteolytic degradation, phosphorylation, and protein–protein interactions. This analysis also reveals the abundance of PEST regions, indicating their importance in eukaryotes.

**ACKNOWLEDGMENTS**

The authors thank Doztányi et al. for providing uninterrupted access to their IUPred server, Dr. S. Ramachandran for bacterial adhesin and nonadhesin sequences, and Eli Eisenberg for housekeeping and tissue-specific gene list. The financial support provided by CSIR as fellowships to G.P.S. and M.G. and project sanctions (CMM0017) to D.D. is duly acknowledged. The authors also thank Dr. Souvik, Chitra, and Dr. Mitali for the valuable suggestions and discussions and Samira for the correction of the manuscript.

**REFERENCES**


