Chapter 2

Intrinsic Disorderness as a Mechanism for PEST Mediated Degradation
2.1 Introduction

PEST regions act as protein degradation signals in many proteins. These regions, rich in proline (P), glutamic acid (E), serine (S) and threonine (T), were first observed in rapidly degraded, eukaryotic intracellular proteins [Rogers et al., 1986]. PEST mediated proteolysis can occur through ubiquitin-proteasome pathway [Rechsteiner and Rogers, 1996; Gregory and Hann, 2000] or by intracellular proteases, calpains [Bordone et al., 2002; Tompa et al., 2004] and caspases [Garay-Malpartida et al., 2005]. Various experimental approaches including deletion, transfer and mutation of PEST regions have shown the role and importance of PEST regions for the stability of proteins [Yaglom et al., 1995; Berset et al., 2002]. However, deletion of PEST regions from certain proteins did not significantly affect their half-life, as PEST are conditional or regulated signals for degradation and altering the cellular environment might allow them to function as degradation signals [Bies et al., 1999; Akgul et al., 2000].

PEST regions have been hypothesized to be solvent exposed because of their enrichment in hydrophilic amino acids [Rechsteiner and Rogers, 1996]. These authors reported conformational flexibility of PEST regions, as many of these regions couldn’t be resolved in X ray structures. The consensus sites of many kinases map to PEST regions and PEST containing proteins (PCPs) are often regulated by reversible phosphorylation [Yaglom et al., 1995; Lin et al., 1996; Marchal et al., 2004] and phosphorylated regions are often disordered [Iakoucheva et al., 2004]. These observations suggest that PEST regions might often be disordered.

Disordered regions are very sensitive to proteolysis as proteases cleave at sterically accessible and flexible sites [Tompa, 2002]. This property has been utilized to determine the location and dynamics of disordered regions in proteins [Iakoucheva et al., 2001; Mark et al., 2005]. Ubiquitination in many proteins also occurs in disordered
regions [Cox et al, 2002] or loop regions [Catic et al, 2004]. Experimental evidence also shows that efficient degradation of polyubiquitinated proteins requires an additional disordered region that serves as the initiation site for degradation and interacts with the proteasomal machinery [Prakash et al, 2004]. Further, 20S proteasome can cleave target proteins at disordered regions independent of ubiquitination [Liu et al, 2003a; Baugh et al, 2009]. These observations show that disordered regions can facilitate rapid degradation of proteins by both proteasome machinery, with or without ubiquitination and by intracellular proteases like calpains and caspases.

If PEST regions are disordered, then this may explain why they act as degradation signals. In the present study I test this hypothesis on experimentally characterized and predicted disorder sequences and show that indeed PEST regions are enriched in disordered regions and proteins. I also show that PEST regions are very common in eukaryotic proteomes and are frequently associated with signaling and transcription regulation function.

2.2 Materials and Methods

2.2.1 Sequence Data


proteins were downloaded from [http://dictybase.org](http://dictybase.org). The list of human housekeeping genes was obtained from [http://www.compugen.co.il/supp_info/Housekeeping_genes.html](http://www.compugen.co.il/supp_info/Housekeeping_genes.html) [Eisenberg and Levanon, 2003]. 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 non-adhesins was obtained from Dr. Ramachandran (personal communication) [Sachdeva et al, 2004].

### 2.2.2 Structure Data

The 4588 nonredundant PDB representative chains were downloaded from the PDB-REPRDB web server [Noguchi and Akiyama, 2003] by applying the following criteria: a) Resolution < 3.0 Å, b) R-factor < 0.3, c) Number of residues > 40, d) % identity ≤ 30 or RMSD > 10 Å.

DSSP (Definition of Secondary Structure Prediction) program [Kabsch and Sander, 1983] was used to calculate the accessible surface area (ASA) of each residue of the PEST region 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, 1994.

### 2.2.3 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](http://www.geneontology.org) and *P. falciparum* GOA file was downloaded from [http://plasmodb.org](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 term present in the respective files. However, there were cases where multiple entries for a single protein with identical GO ID were
observed. Only unique entries were considered for calculating the over/under representation of GO terms in PCPs.

### 2.2.4 Prediction of PEST Regions

The EMBOSS program *epestfind* was used to find PEST regions in 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 region. The quality of a PEST region is determined by a scoring parameter based on the local enrichment of the critical amino acids and its hydrophobicity. A score of ≥ +5.0 obtained for a region with the *epestfind* is taken as a valid PEST motif. For all the analysis, I only considered valid PEST regions.

### 2.2.5 Disorder Prediction

Locally installed DisEMBL “Hot Loop” prediction with default parameters (http://dis.embl.de) was used for local disorder prediction in the human PCPs [Linding et al, 2003].

The modified Uversky’s method earlier developed by our group was applied on the complete eukaryotic proteomes to predict disordered and ordered proteins [Pandey et al, 2004]. Modified Uversky method is based on the observation that IDPs have high mean net charge and/or low mean hydrophobicity [Uversky et al, 2000]. It is a global disorder predictor and gives a Netscore to each protein. This was calculated for each protein using following formula:

\[
\text{Netscore} = \left( \frac{\text{mean net charge}}{2.785 \times \text{mean net hydrophobicity} + 1.151} \right) \div 2.952
\]

A protein with positive Netscore is predicted to be IDP, while negative score predicts ordered protein.
IUPred program was also used for protein disorder prediction in the eukaryotic proteins [Dosztanyi et al, 2005a; Dosztanyi et al, 2005b]. Proteins with ≥ 50% of the residues having a score of ≥ 0.5 were denoted as IDPs and the rest were categorized as ordered.

### 2.2.6 N/C Terminal of Proteins

The initial 25% and the final 25% of the amino acid sequence of a protein were denoted as N and C terminals respectively [Prakash et al, 2005].

### 2.2.6 Statistics

Chi-square test was applied for calculating the differences in the distribution of PCPs in 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 [Crooks et al, 2004] was used to construct Pro-X-Pro-X-Pro pattern from all predicted human PEST regions.

### 2.3 Results

#### 2.3.1 Prevalence of PEST Regions in Characterized Ordered and Disordered Sequences

PEST regions have been hypothesized to be disordered as they are enriched in disorder promoting amino acid residues [Wright and Dyson, 1999; Tompa, 2002]. To analyze whether PEST are enriched in experimentally characterized disorder regions, I examined a dataset consisting of 146 disordered sequences and 1191 globular segments as described in the ‘Sequence Data’ section of ‘Materials and Methods’. Only 3.03% of globular protein segments contained PEST regions in comparison to 17.81% of disordered sequences (Table 2.1). This was significant over-representation of PEST regions in disordered sequences in comparison to globular ones ($\chi^2 = 64.294$, df = 1, $P$-value = 1.07E-15).
Table 2.1 Distribution of PEST regions in disordered and globular/ordered sequences.

<table>
<thead>
<tr>
<th>Category</th>
<th>Total no. of sequences</th>
<th>No. of sequences containing PEST [%]</th>
<th>Total no. of PEST regions#</th>
</tr>
</thead>
<tbody>
<tr>
<td>Disordered</td>
<td>146</td>
<td>26 [18]</td>
<td>47</td>
</tr>
<tr>
<td>sequences</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Globular</td>
<td>1191</td>
<td>36 [3]</td>
<td>38</td>
</tr>
<tr>
<td>sequences</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

# Total number of PEST regions present in all the disordered/ globular regions.

2.3.2 Distribution of PEST Regions in Predicted Ordered and Disordered proteins

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

I further analyzed the distribution of PEST regions in eukaryotic predicted disordered and ordered proteins using the modified Uversky method [Pandey et al, 2004] and IUPred program [Dosztanyi et al, 2005a; Dosztanyi et al, 2005b] (see “Materials and Methods”, for definition of disordered proteins). The results of both methods show that predicted disordered proteins are enriched in PEST regions (Figure 2.1). In all proteomes analyzed, a highly significant difference in the distribution of PCPs was observed between predicted disordered and ordered proteins.
Figure 2.1 Distribution of PEST regions in eukaryotic predicted IDPs and ordered proteins (see “Materials and Methods” for definition of IDPs and ordered proteins).
A. The modified Uversky method and B. IUPred program were used. The results of both methods show that predicted IDPs are significantly enriched in PEST regions as compared to ordered proteins. The chi-square $P$-values for the differential distribution of PEST containing proteins in predicted IDPs and ordered proteins are tabulated in Supplementary Table I.

Also, the fraction of residues predicted by IUPred was substantially higher in PEST regions compared to the whole human proteome (46 % vs. 26 % respectively, Figure 2.2).

![Graph showing higher fraction of disordered residues in PEST regions.]

**2.3.3 PEST Regions in PDB Structures**

Analysis of 4588 unique PDB representative chains (see “Materials and Methods”) show 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 significant number of residues were unstructured (69% irregular [U], 14% β-sheet [E] and 17% α-helix [H]) as compared to non-PEST regions (48% irregular [U], 20% β-sheet [E] and 32% α-helix [H]), Figure 2.3.
Figure 2.3 Proportion of residues in different secondary structure classes in PEST and non-PEST regions in PDB.

The surface accessibility analysis for PEST regions was carried out on 179 (i.e., ≥ 50% resolved) PEST regions, out of which 149 were more than 50% surface exposed (Figure 2.4).
Intrinsic Disorderness as a Mechanism for PEST Mediated Degradation

**Figure 2.4** Surface accessibility of PEST residues in the representative PDB chains. 149 out of 179 PEST regions (≥ 50% resolved) were found to be more than 50% surface exposed.

### 2.3.4 Abundance of PCPs in Eukaryotes

Analysis of 11 completely sequenced eukaryotes show that PCPs make up a substantial fraction of the proteome (~25%) (Figure 2.5). *P. falciparum* and *E. cuniculi* were exceptions as they have low fraction of PCPs (~11%). This might not be related to their intracellular parasite status, as proteins of partially sequenced *L. major* and *C. parvum* showed a general trend (28% and 21% respectively) of PCP distribution. The extremely high AT% of *P. falciparum* genome may also not be related to this observation, as other AT rich organisms such as *D. discoideum* and *C. parvum* followed the general trend.
Figure 2.5 Prevalence of PCPs in completely sequenced eukaryotes. Except for the low representation in *E. cuniculi* and *P. falciparum*, other eukaryotes have about 25% of the proteome as PCPs.

### 2.3.5 Functional Classification of PEST Containing Proteins

Classification of eukaryotic PCPs into Gene Ontology classes indicated their over-representation in categories related to transcription regulation, signaling, endocytosis, and under-representation in metabolism and transport proteins (Supplementary Table II). PCPs were also over-represented in human non-housekeeping as compared to housekeeping proteins (37% vs. 28% respectively, $\chi^2 = 16.301$, df = 1, *P*-value = 0.0001, Figure 2.6).
The classification of *P. falciparum* PCPs revealed that about one third of all PEST regions were present in 54 PfEMP-1 family of extracellular proteins (Supplementary Table II). ‘Cysteine type proteases’ was the only intracellular class that showed significant over-representation of PEST regions in *P. falciparum*.

### 2.3.6 Positional Preference of PEST Regions in Proteins

PEST regions have been reported to be preferentially present in the C terminal region of the proteins [Rechsteiner and Rogers, 1996]. Although both N and C terminal PEST regions have been experimentally characterized [Medintz *et al*, 2000; Schoonbroodt *et al*, 2000], I observed that reports of C terminal PESTs are more prevalent. To analyze for any preferential localization of PEST regions within a protein, I mapped 19,071 PEST regions in 9,827 human PCPs. 4,396 PEST regions were mapped to the N terminals and 4,638 to the C terminals. There was no preferential localization of PEST regions in the C terminals of proteins. I also analyzed for differences in scores of PEST regions in the N and C terminals regions.
$P$-value of the two tailed, unpaired t-test for this analysis was found to be insignificant ($P = 0.1$).

2.4 Discussion

PEST regions have been extensively studied as protein degradation signals and their role as phosphorylation targets and protein–protein interaction sites are also reported [Yaglom et al., 1995; Lin et al., 1996; Rechsteiner and Rogers, 1996; Van Antwerp and Verma, 1996; Marchal et al., 1998]. These regions are rich in hydrophilic amino acids and have been hypothesized to be disordered. In this study, I have analyzed both the disorderness of PEST regions as well as their distribution in the eukaryotic proteomes.

2.4.1 Disorderness of PEST Regions Can Explain Why They Act as Degradation Signals

My data reveals that PEST regions are over-represented in characterized, non-redundant disordered sequences (Table 2.1). Predicted IDPs in eukaryotes also show a significant enrichment of PEST regions as compared to the ordered proteins (Figure 2.1). The low occurrence of PEST regions in representative PDB chains indicates that these regions might be predominantly disordered because disordered regions are often removed prior to crystallization, as they interfere with the crystallization process. The prevalence of irregular secondary structure (Figure 2.2) and surface accessible residues (Figure 2.3) in the PEST regions in PDB further indicates their disorderness. Enrichment of a proline-rich motif and a charged pattern has been reported in disordered regions [Lise and Jones, 2005]. I examined the Pro-X-Pro-X-Pro motif composition in all the predicted human PEST regions using WebLogo server [Crooks et al., 2004]. The composition of proline rich motif Pro-X-Pro-X-Pro in PEST regions was similar to the pattern seen in disordered regions (Supplementary Figure 1). Taken together, these data clearly indicate the disorderness of PEST regions in eukaryotes.
Disordered regions have been shown to be involved in intracellular protein degradation by a variety of mechanisms. These are highly sensitive to proteases in vitro and possibly also to intracellular proteases [Tompa, 2002]. Ubiquitination within disordered regions has been reported for many proteins [Cox et al, 2002]. Structural analysis of preferred ubiquitination sites also revealed preference for loop regions [Catic et al, 2004]. It has been shown that efficient degradation of polyubiquitinated proteins requires an additional disordered region that serves as the initiation site for degradation and interacts with the proteasomal machinery [Prakash et al, 2004]. Degradation of non-ubiquitinated proteins by proteasome within disordered regions has also been reported [Liu et al, 2003a; Baugh et al, 2009]. Thus disordered regions can mediate rapid degradation of target proteins.

Even though PEST regions have been studied extensively, why do they act as degradation signals has not been clear. My data that PEST regions are often disordered combined with the information that disordered regions are implicated in degradation of proteins, can explain why PEST regions act as protein degradation signals. This also explains the observation that PEST regions are often phosphorylated and are involved in protein binding [Rechsteiner and Rogers, 1996; Lin et al, 1996]. It has recently been observed that disordered proteins have significantly higher protein degradation rates [Tompa et al, 2007; Gsponer et al, 2008]. This may at least partially be due to enrichment of PEST regions in the disordered proteins. Gsponer et al, 2009, find higher degradation rates for mRNAs encoding disordered proteins and also their lower noise levels. Thus it seems disordered proteins are regulated at different levels to ensure their tight regulation [Gsponer et al, 2008]

2.4.2 Functions of PCPs in Eukaryotic Proteomes

My proteome-wide data shows 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 classes like nucleic acid and protein binding, transcriptional regulation and signal transduction might reflect increased requirement of regulated proteolysis in these classes. Earlier studies have also reported the enrichment of PEST regions in various regulatory classes of proteins including phosphatases, annexins, cyclic nucleotide signaling pathway proteins, EF-hand calcium-binding proteins and calmodulin binding proteins [Barnes and Gomes, 1995; Gomes and Barnes, 1995; Gomes and Barnes, 1997; Sekhar and Freeman, 1998; Barnes and Gomes, 2002].

The low incidence of PCPs in \textit{P. falciparum} and \textit{E. cuniculi} may indicate less dependence of these organisms on PEST-mediated proteolysis or involvement of an altered PEST motif. Analysis of \textit{P. falciparum} proteins revealed enrichment of PEST regions in PfEMP-1 class of extracellular proteins. These proteins are transported to the membrane of the infected red blood cells and mediate their adhesion and clumping on vascular endothelium cells [Kyes \textit{et al}, 2001]. The cell adhesion caused by these proteins leads to most malaria related deaths. An earlier study on chromosome 2 proteins of \textit{P. falciparum} also found abundance of PEST regions in this class of proteins [Mitchell and Bell, 2003]. On the contrary, Rifins - another class of host cell membrane targeted proteins with unclear function, were devoid of PEST regions. I also found over-representation of PEST regions in bacterial cell adhesion proteins as compared to non-adhesins (18\% as compared to 11\%, $\chi^2 = 12.611$, df = 1, $P$-value = 0.0004, Figure 2.7). These data indicates a role of PEST-like sequences in cell adhesion processes.
‘Cysteine-type proteases’ was the only intracellular class in *P. falciparum* with over-representation of PCPs. Particularly 6 of the 8 SERA (Serine Repeat Antigen) proteins contains 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 [Li *et al*, 2002]. PEST regions may play a role in the proteolytic processing and regulation of intracellular levels of SERA proteins.

In a previous study, a small dataset of 8 proteins containing functional PEST regions was analyzed and significant preference of PEST regions in carboxy terminal regions was found [Rechsteiner and Rogers, 1996]. My proteome-wide data doesn’t reflect any bias towards the presence of predicted PEST regions in the carboxy terminals of proteins.