Published in final edited form as:
Science. 2008 November 28; 322(5906): 1365–1368. doi:10.1126/science.1163581.

Tight regulation of unstructured proteins:

from transcript synthesis to protein degradation

Jörg Gsponer[1,*], **Matthias E. Futschik**[1,2,3], **Sarah A. Teichmann**[1], and **M. Madan Babu**[1,*]

[1]MRC Laboratory of Molecular Biology, Hills Road, Cambridge CB2 0QH, United Kingdom

[2]Institute for Theoretical Biology, Charité, Humboldt-University, Berlin, Germany

[3]Centre of Molecular and Structural Biomedicine, University of Algarve, Faro, Portugal

Abstract

Altered abundance of several intrinsically unstructured proteins (IUPs) has been associated with perturbed cellular signalling that may lead to pathological conditions such as cancer. Therefore, it is important to understand how cells precisely regulate availability of IUPs. We observe that regulation of transcript clearance, proteolytic degradation, and translational rate contribute to control the abundance of IUPs, some of which are present in low amounts and for short periods of time. Abundant phosphorylation and low stochasticity in transcription and translation indicates that availability of IUPs can be finely tuned. Fidelity in signalling may require that most IUPs are available in appropriate amounts and not present longer than needed.

Up to one-third of all eukaryotic proteins have large segments that are unstructured and are commonly referred as intrinsically unstructured proteins (IUPs). These proteins lack a unique structure, either entirely or in parts, when alone in solution (1). The lack of structure is thought to provide several advantages such as (i) increased interaction surface area, (ii) conformational flexibility to interact with several targets, (iii) presence of molecular recognition elements that fold upon binding, (iv) accessible post-translational modification sites and (v) availability of short linear interaction motifs (2-5). These and other properties are ideal for proteins that mediate signaling and coordinate regulatory events and indeed proteins that participate in regulatory and signaling functions are enriched in unstructured segments (6-8) (SOM text S1). Because of their unusual structural and important functional properties, the presence of IUPs in a cell may need to be carefully monitored. In fact, altered abundance of IUPs is associated with several disease conditions. For instance, over-expression of TC-1 (9) or under-expression of Arf (10) and p27 (11) has been linked with various types of cancer. Similarly, over-expression of α-synuclein and tau increases the risk of aggregate formation and has been linked to Parkinson's disease and Alzheimer's disease (12, 13). We therefore tested whether specific control mechanisms affect the availability of IUPs (that is their abundance and residence time) within a cell.

Using the Disopred2 software (6), which reports unstructured residues based on the protein sequence, we computed the fraction of the polypeptide that is predicted to be unstructured for every protein in *Saccharomyces cerevisiae* (14). This allowed us to categorize 1971 sequences as highly structured (S: 0 to 10% of the total length is unstructured), 2711 sequences as moderately unstructured (M: 10 to 30% of the protein is unstructured) and 2020 sequences as highly unstructured (U: 30 to 100% of the protein is unstructured) (Fig. 1). This information was integrated with different genome-scale datasets describing most of the regulatory steps that influence protein synthesis or degradation (table S1, fig. S1) and the

*To whom correspondence should be addressed. Emails: MMB (madanm@mrc-lmb.cam.ac.uk) or JG (jgsponer@mrc-lmb.cam.ac.uk).

Europe PMC Funders Author Manuscripts

distributions of the values for the proteins in the highly unstructured (U) and highly structured groups (S) were compared using statistical tests (14).

We compared transcription of genes encoding highly unstructured to that of genes encoding more structured proteins. Because the steady state amount of mRNA could be affected by the rate at which the transcripts are produced or degraded, we investigated whether the transcriptional rate or the degradation rate were different for the transcripts that encode highly structured and unstructured proteins. The number of transcription factors (TFs) that regulate a gene was comparable between the two groups (p=0.55; Wilcoxon test; Fig. 2A). However, mRNAs encoding highly unstructured proteins were generally less abundant than transcripts encoding more structured proteins ($p=1\times10^{-6}$; Wilcoxon test; Fig. 2B). The mRNA half-lives of the transcripts encoding highly unstructured proteins were lower than transcripts encoding more structured proteins ($p<10^{-16}$; Wilcoxon test; Fig. 2C) and a comparison of the distribution of transcriptional rates showed that the difference between the two groups was less significant ($p=1\times10^{-8}$; Wilcoxon test; table S3). Thus, differences in decay rates appear to be a major factor leading to differences in mRNA abundance (SOM text S5).

We analysed poly-A tail length because the two major pathways of mRNA decay are initiated by removal of the poly-A tail. A significantly larger fraction of the unstructured proteins had a short poly-A tail (table S1) than did structured proteins ($p<10^{-16}$; Fisher's exact test; Fig. 2D). Analysis of transcript binding by Puf family RNA-binding proteins, which affect transcript stability, showed that Puf5p binding was enriched for transcripts that encode highly unstructured proteins. In fact, 108 of the 224 transcripts bound by Puf5p encode highly unstructured proteins, a much greater number than expected by chance (expected: 68 transcripts; z-score: 5.3; $p=5\times10^{-10}$; (14)). Thus poly-A tail length and interaction with specific RNA-binding proteins may modulate stability of transcripts encoding IUPs (SOM text S5).

Unstructured proteins tend to be less abundant than structured proteins ($p<10^{-16}$; Wilcoxon test; Fig. 2F, fig. S2; SOM text S2, S5). The rate of protein synthesis was significantly lower ($p<10^{-16}$; Wilcoxon test; Fig. 2E) and protein half-life was shorter ($p=1\times10^{-15}$; Wilcoxon test; Fig 2G; SOM text S5) for highly unstructured than for more structured proteins. Two pathways that mediate ubquitin-proteasome-dependent degradation are the N-end rule pathway and PEST mediated degradation pathway. Although the distribution of N-end residues was not significantly different (SOM text S3; figs. S3, S4), a significantly greater fraction of the unstructured proteins contained PEST motifs ($p<10^{-16}$; Fisher's exact test; Fig. 2H; SOM text S5) as previously observed (1, 15). Therefore, it appears that the availability of many IUPs is regulated via proteolytic degradation and a reduced translational rate.

For certain IUPs (*e.g.*, p27), post-translational modifications such as phosphorylation (11, 16) can affect their abundance or half-life in a cell. In fact, recent computational studies using phosphorylation site prediction methods have suggested that unstructured regions are enriched for sites that can be post-translationally modified (17). We analyzed the experimentally determined yeast kinase-substrate network and found that highly unstructured proteins are on average substrates of twice as many kinases as are structured proteins ($p=1\times10^{-12}$; Wilcoxon test; SOM text S4; fig. S5). Notably, on average, 51±19% (S.D.) of all substrates of the kinases are highly unstructured whereas only 19±13% (S.D.) are highly structured (the remaining 30±14% (S.D.) of all substrates are moderately unstructured). This is a significant bias compared to the expected genome-wide distribution of ~30% highly unstructured and ~30% structured proteins based on our categorization ($p<10^{-16}$; Fisher's exact test (14)). We found that 85% of the kinases for which more than

50% of their substrates are highly unstructured (table S2) are either regulated in a cell-cycle dependent manner (*e.g.*, Cdc28), or activated upon exposure to particular stimuli (*e.g.*, Fus3) or stress (*e.g.*, Atg1). This suggests that post-translational modification of IUPs through phosphorylation may be an important mechanism in fine-tuning their function and possibly their availability under different conditions.

We investigated whether genes that encode highly unstructured proteins display low stochasticity in their expression levels among individuals in a population of genetically homogeneous cells. An important source of such stochasticity in cellular systems is random noise in transcription and translation, which results in very different amounts of transcripts and protein products. We used the presence of a TATA box in the promoter region to infer genes that might be more subjected to noise in gene expression (18) and found that genes encoding highly unstructured proteins are less likely to have a TATA box than those that encode structured proteins ($p=8\times10^{-7}$; Fisher's exact test; Fig. 2I). At the protein level, analysis of direct experimental data revealed that unstructured proteins have lower noise levels, when compared to structured proteins ($p=3\times10^{-11}$; Wilcoxon test; Fig. 2J). These results indicate that highly unstructured proteins may display less noise in transcription and translation than more structured proteins.

To assess regulation of IUPs in other organisms, we analyzed datasets (table S1) for *Schizosaccharomyces pombe* and *Homo sapiens*. Similar trends to those observed for *S. cerevisiae* were evident in these organisms (Fig. 3; tables S3, S4). Thus both unicellular and multicellular organisms appear to regulate the availability of IUPs. The observed differences between structured and unstructured proteins were independent of the IUP prediction method used, protein length, localisation within the major sub-cellular compartments, different grouping of proteins or the number of interaction partners per protein (tables S6 to S11). Though the distributions of the values for the proteins in the structured and unstructured groups are broad and overlap, the differences reported here are consistently statistically significant using two independent statistical tests: Wilcoxon rank-sum test and Kolmogorov-Smirnov test (table S3 to S5, (14)). Thus the reported trends appear to be attributable to the intrinsically unstructured nature of the proteins. Intriguingly, of all the IUPs, those that contain poly-Q or poly-N stretches seem to be more tightly controlled (table S3, S4).

Proteins with unstructured regions predominantly have signalling or regulatory roles and are often re-used in multiple pathways to produce different physiological outcomes (2, 6-8). Accordingly, increased abundance of IUPs can result in undesirable interactions (for example, titration of unrelated proteins by inappropriate interaction through exposed peptide motifs), thereby disturbing the fine balance of the signalling networks leading to dampened or inappropriate activities (19). Spatial and temporal segregation of signalling proteins as well as an increased signalling complexity may contribute to fidelity in regulation (SOM text S5, fig. S6). In addition, tight regulation of signalling and regulatory IUPs could minimize the potentially harmful effects of ectopic interactions and may provide robustness to signalling processes by ensuring that such proteins are present in appropriate amounts and time periods. Indeed, free IkB (an IUP that interacts with and inhibits the NF-kB transcription factor) must be continuously degraded to allow for rapid and robust NF-kB activation and to make the pathway sensitive for a signalling input (20, 21). In contrast, stabilization of free IkB by removal of the PEST motif reduces NF-kB activation (20, 21). In mammalian cells exposed to mild ER stress, survival is favoured as a direct consequence of the intrinsic instability of mRNAs and of proteins that drive apoptosis (such as Chop, an IUP) compared to that of proteins that promote adaptation and cell longevity (for example the chaperone protein BiP, a structured protein) (22).

Although the abundance of many IUPs is strictly controlled, certain IUPs are present in cells in large amounts or for long periods of time. In fact, in some cases (for example the fibrous muscle protein, titin), large amounts may be required throughout the life-time of a cell. Fine-tuning of IUP availability can be achieved by post-translational modifications and interactions with other factors (1, 3, 11, 16). Both mechanisms can promote increased abundance or longer half-life through changes in cellular localization or by protection from the degradation machinery (*e.g.*, certain phosphorylated forms of the cyclin dependent kinase inhibitory protein $p27^{kip1}$ and the spinocerebellar ataxia type 1 protein ataxin-1) (23, 24). Although association with other proteins may increase their stability, free IUPs are likely to be rapidly degraded by the 20S proteasome via degradation by default (25) as shown for the unbound forms of $p21^{cip1}$ (26), $p27^{kip1}$ (27), α-synuclein (28) and tau (29). Certain post-translational modifications may promote regulated degradation (for example that of $p27^{Kip1}$) (11, 16). In this context, our finding that many IUPs tend to be phosphorylated by multiple kinases and display low noise in transcription and translation suggests that their abundance and half-lives may be finely tuned in cells (SOM text S5).

Our studies reveal an evolutionarily conserved tight control of synthesis and clearance of most IUPs. The discovery was made possible by integrating multiple large-scale datasets describing control mechanisms during transcription, translation, and post-translational modification with structural information on proteins. Besides the elucidation of general trends, the framework describing multiple levels of regulation introduced here, might facilitate investigation of how individual IUPs are fine-tuned in different cell types and how perturbations to this tight control might influence disease conditions (30).

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The authors acknowledge the MRC for funding and would like to thank the two anonymous referees, the editor, A Bertolotti, A Tzakos, A Wuster, C Brockmann, C Chothia, D Finley, D Rubinsztein, E Levy, H McMahon, I Schafer, J Bui, K Weber, M Goedert, R Janky, S Michnick and S Sarkar for providing helpful comments. We apologise for not citing the work describing the datasets and other references due to lack of space. Many can be found in the SOM. MMB acknowledges Darwin College and Schlumberger Ltd, MMB and MF acknowledge the Royal Society for support. JG is funded by an MRC Special Training Fellowship in Computational Biology.

References

1. Dunker AK, et al. J Mol Graph Model. 2001; 19:26. [PubMed: 11381529]
2. Wright PE, Dyson HJ. J Mol Biol. 1999; 293:321. [PubMed: 10550212]
3. Kriwacki RW, Hengst L, Tennant L, Reed SI, Wright PE. Proc Natl Acad Sci U S A. 1996; 93:11504. [PubMed: 8876165]
4. Tompa P. FEBS Lett. 2005; 579:3346. [PubMed: 15943980]
5. Oldfield CJ, et al. Biochemistry. 2005; 44:12454. [PubMed: 16156658]
6. Ward JJ, Sodhi JS, McGuffin LJ, Buxton BF, Jones DT. J Mol Biol. 2004; 337:635. [PubMed: 15019783]
7. Liu J, et al. Biochemistry. 2006; 45:6873. [PubMed: 16734424]
8. Iakoucheva LM, Brown CJ, Lawson JD, Obradovic Z, Dunker AK. J Mol Biol. 2002; 323:573. [PubMed: 12381310]
9. Sunde M, et al. Cancer Res. 2004; 64:2766. [PubMed: 15087392]
10. Sherr CJ. Nat Rev Cancer. 2006; 6:663. [PubMed: 16915296]
11. Grimmler M, et al. Cell. 2007; 128:269. [PubMed: 17254966]
12. Chiti F, Dobson CM. Annu Rev Biochem. 2006; 75:333. [PubMed: 16756495]

13. Goedert M. Nat Rev Neurosci. 2001; 2:492. [PubMed: 11433374]
14. (Materials and methods are available as supporting material on *Science* Online).
15. Tompa P, Prilusky J, Silman I, Sussman JL. Proteins. 2007
16. Chu I, et al. Cell. 2007; 128:281. [PubMed: 17254967]
17. Iakoucheva LM, et al. Nucleic Acids Res. 2004; 32:1037. [PubMed: 14960716]
18. Raser JM, O'Shea EK. Science. 2004; 304:1811. [PubMed: 15166317]
19. Pawson T, Warner N. Oncogene. 2007; 26:1268. [PubMed: 17322911]
20. Mathes E, O'Dea EL, Hoffmann A, Ghosh G. Embo J. 2008; 27:1421.
21. O'Dea EL, et al. Mol Syst Biol. 2007; 3:111. [PubMed: 17486138]
22. Rutkowski DT, et al. PLoS Biol. 2006; 4:e374. [PubMed: 17090218]
23. Chen HK, et al. Cell. 2003; 113:457. [PubMed: 12757707]
24. Chu IM, Hengst L, Slingerland JM. Nat Rev Cancer. 2008; 8:253. [PubMed: 18354415]
25. Asher G, Reuven N, Shaul Y. Bioessays. 2006; 28:844. [PubMed: 16927316]
26. Li X, et al. Mol Cell. 2007; 26:831. [PubMed: 17588518]
27. Tambyrajah WS, Bowler LD, Medina-Palazon C, Sinclair AJ. Arch Biochem Biophys. 2007; 466:186. [PubMed: 17854759]
28. Tofaris GK, Layfield R, Spillantini MG. FEBS Lett. 2001; 509:22. [PubMed: 11734199]
29. David DC, et al. J Neurochem. 2002; 83:176. [PubMed: 12358741]
30. Balch WE, Morimoto RI, Dillin A, Kelly JW. Science. 2008; 319:916. [PubMed: 18276881]

Europe PMC Funders Author Manuscripts

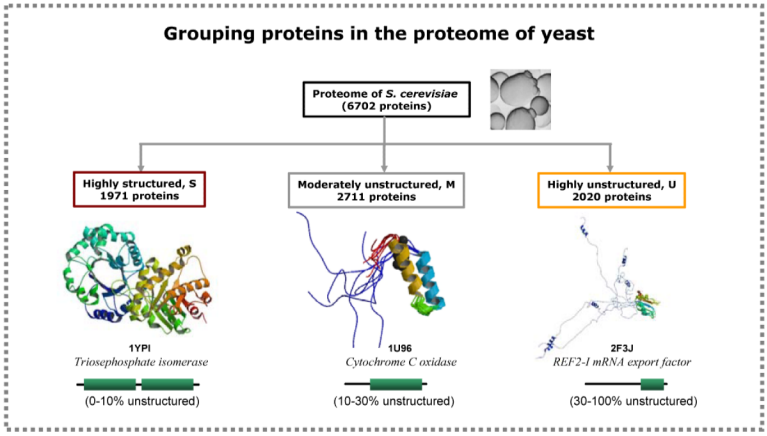

Fig. 1.

The *S. cerevisiae* proteome was grouped into three categories, the highly structured (S), moderately unstructured (M) and the highly unstructured (U) category based on the fraction of unstructured residues over the entire protein length.

Europe PMC Funders Author Manuscripts

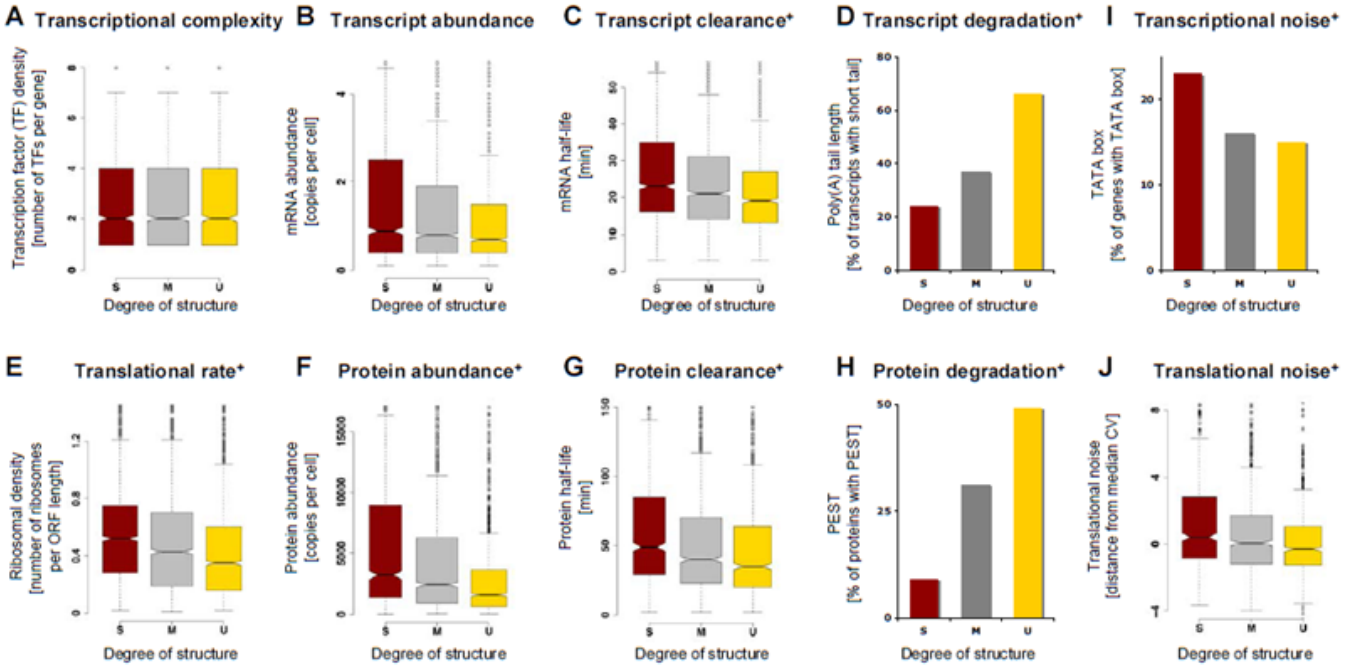

Fig. 2.

Box-plots for various regulatory and cellular properties for the three different groups of proteins (S, M and U) in *S. cerevisiae*. Each box-plot identifies the middle 50% of the data, the median, and the extreme points. The entire set of datapoints is divided into quartiles and the interquartile range (IQR) is calculated as the difference between $x_{0.75}$ and $x_{0.25}$. The range of the 25% of the data points above ($x_{0.75}$) and below ($x_{0.25}$) the median ($x_{0.50}$) is displayed as a filled box. The horizontal line and the notch represent the median and confidence intervals, respectively. Datapoints greater or less than 1.5*IQR represent outliers and are shown as dots. The horizontal line that is connected by dashed lines above and below the filled box (whiskers) represents the largest and the smallest non-outlier datapoints, respectively. (**A**) transcriptional complexity, (**B**) mRNA abundance, (**C**) mRNA half-life, (**D**) poly-A tail length (percentage of transcripts that have short poly-A tail length within a group), (**E**) ribosomal density, (**F**) protein abundance, (**G**) protein half-life, (**H**) PEST-sequence content (percentage of proteins with PEST motif within a group), (**I**) TATA box content (percentage of genes with a TATA box within a group) and (**J**) noise (distance from median co-efficient of variation) in protein production. $^{+}$ denotes statistically significant differences between the S and U group (table S3).

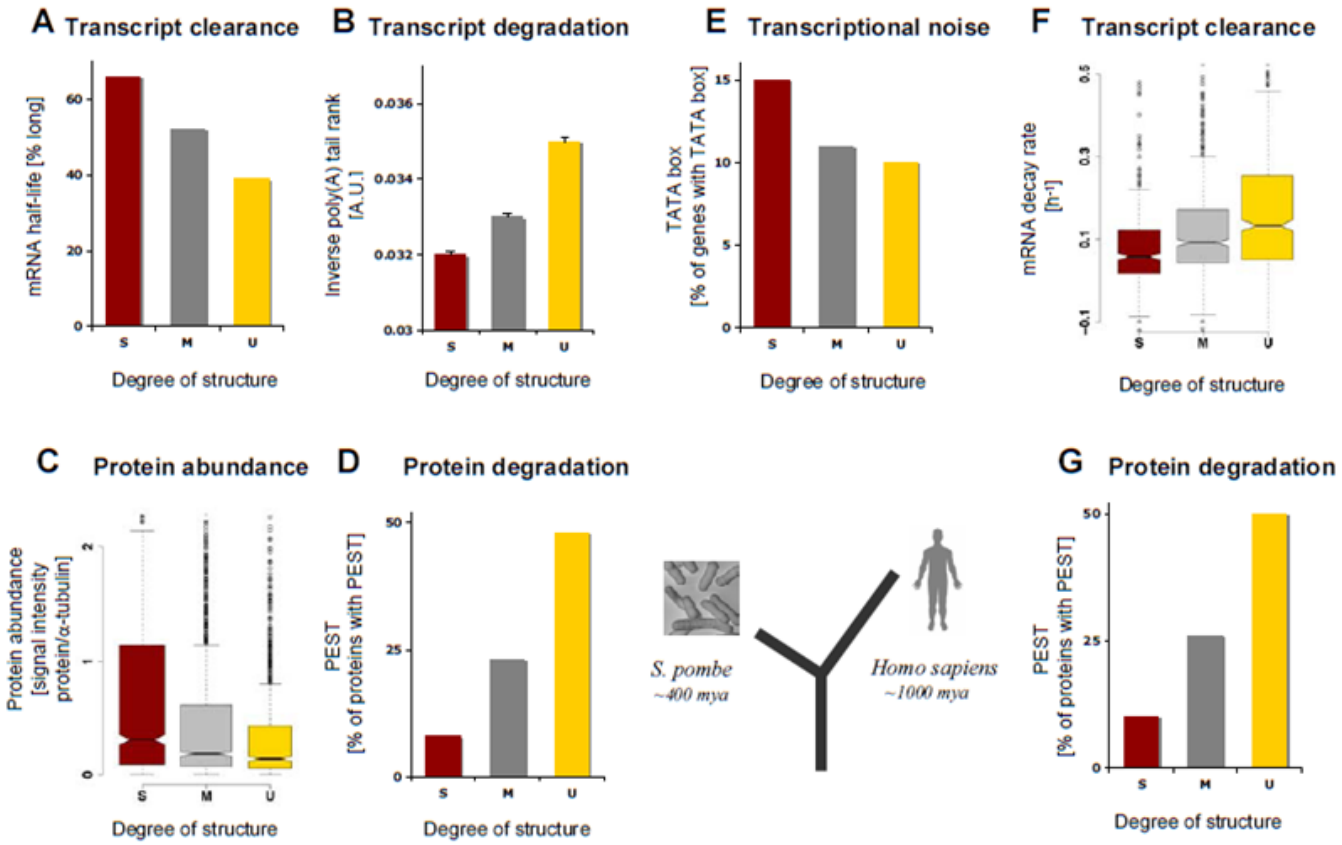

Fig. 3.

Box-plots for various regulatory and cellular parameters for the three groups of proteins (S, M and U) in *S. pombe* (**A-D**) and humans (**E-G**). All reported differences are statistically significant (table S4).

Europe PMC Funders Author Manuscripts