The following resources related to this article are available online at
www.sciencemag.org (this information is current as of November 28, 2008):

Updated information and services, including high-resolution figures, can be found in the online
version of this article at:
http://www.sciencemag.org/cgi/content/full/322/5906/1365

Supporting Online Material can be found at:
http://www.sciencemag.org/cgi/content/full/322/5906/1365/DC1

This article **cites 29 articles**, 5 of which can be accessed for free:
http://www.sciencemag.org/cgi/content/full/322/5906/1365#otherarticles

This article appears in the following **subject collections**:
Biochemistry
http://www.sciencemag.org/cgi/collection/biochem

Information about obtaining **reprints** of this article or about obtaining **permission to reproduce
this article** in whole or in part can be found at:
http://www.sciencemag.org/about/permissions.dtl

in the direction perpendicular to the ridge axis on the east side of the ridge at depths between 70 and 120 km (*4*, *5*). Such structure was not observed on the west side of the axis (*4*, *5*). Both conductive structures show strong anisotropy that cannot be accounted for by hydrated olivine crystals (*15*). An additional conductive component is therefore needed, and the presence of silicate melt has been suggested (*15*).

Figure 2 shows in a conductivity-depth plot the different electrical structures of the Pacific Ocean mantle described above. We also show electrical conductivities calculated for different model mantles consisting of dry and hydrous olivine, dry olivine with 5 to 30 volume % of MORB and dry olivine with variable amounts of carbonatite melt. All calculated conductivities vary with depth following an adiabatic path (*19*, *20*). The deep conductive asthenosphere (70 to 200 km) under the ridge axis can be explained considering olivine mixed with between 0.035 and 0.35 volume % of carbonatite. Olivine with more than 5 volume % of basalt could alternatively account for the high conductivity of this region, but an elevated degree of peridotite melting is precluded at such depth (*11*, *19*). Hydrous olivine is unable to account for such a high conductivity domain (*14*, *15*). The presence of a small amount of carbonatite is therefore the best explanation for the high conductivity of this region. Furthermore, the range 0.035 to 0.35 volume % of carbonatite corresponds (*31*) on average to ~300 ppmw of CO_2 stored in the asthenosphere, in reasonable agreement with the variable estimates of CO_2 concentrations in the source region of MORBs (*16–18*, *32*). The off-axis conductive asthenosphere (depths of 70 to 120 km in Fig.2) can not be explained by olivine with a few percent of basalt, because it extends at depths inconsistent with silicate melting (*11*). Hydrous olivine may account for its lowermost conductivity but not for the upper range of values. From our results, the off-axis asthenosphere can be interpreted as olivine with ~0.005 to 0.035 volume % of carbonatite melts.

It has been shown that the off-axis conductive zone joins the deep asthenosphere under the ridge (*5*). Carbonatite melts under the ridge derive from the decompression melting initiating at depths of 300 km (*20*), whereas their presence extending 350 km eastward of the ridge at depths of 70 to 120 km might result from their lateral migration. Based on conductivities shown in Fig. 2, we suggest that about one-tenth of the carbonatite melt produced in the asthenosphere under the ridge may flow laterally in the mantle. The rest would migrate upward and be incorporated in the silicate melts deriving from peridotite melting at depths ~70 to 80 km below the ridge. The CO_2 contents in MORBs and their heterogeneity (*16–18*) would therefore originate from the incorporation of small amount of carbonatite melts by basaltic melts.

References and Notes

1. Y. Xu, T. J. Shankland, B. T. Poe, *J. Geophys. Res.* **105**, 27865 (2000).
2. S. Constable, *Geophys. J. Int.* **166**, 435 (2006).
3. A. G. Jones, *Lithos* **48**, 57 (1999).
4. R. L. Evans *et al.*, *Nature* **437**, 249 (2005).
5. K. Baba *et al.*, *J. Geophys. Res.* **111**, B02101 (2006).
6. F. Gaillard, G. I. Marziano, *J. Geophys. Res.* **110**, B06204 (2005).
7. J. A. Tyburczy, H. S. Waff, *J. Geophys. Res.* **88**, 2413 (1983).
8. S.-I. Karato, in *Water in Nominally Anhydrous Minerals, Book Series: Reviews in Mineralogy and Geochemistry*, H. Keppler, J. R. Smyth, Eds. (Mineralogical Society of America, Chantilly, VA, 2006), vol. 62, chap.15.
9. F. Gaillard, *Earth Planet. Sci. Lett.* **218**, 215 (2004).
10. A. Pommier, F. Gaillard, M. Pichavant, B. Scaillet, *J. Geophys. Res.* **113**, B05205 (2008).
11. C. Aubaud, E. H. Hauri, M. M. Hirschmann, *Geophys. Res. Lett.* **31**, L20611 (2004).
12. P. Tarits, S. Hautot, F. Perrier, *Geophys. Res. Lett.* **31**, L06612 (2004).
13. S. Demouchy, S. D. Jacobsen, F. Gaillard, C. R. Stern, *Geology* **34**, 429 (2006).
14. D. J. Wang, M. Mookherjee, Y. S. Xu, S.-I. Karato, *Nature* **443**, 977 (2006).
15. T. Yoshino, T. Matsuzaki, S. Yamashita, T. Katsura, *Nature* **443**, 973 (2006).
16. A. E. Saal, E. H. Hauri, C. H. Langmuir, M. R. Perfit, *Nature* **419**, 451 (2002).
17. P. Cartigny, F. Pineau, C. Aubaud, M. Javoy, *Earth Planet. Sci. Lett.* **265**, 672 (2008).
18. B. Marty, I. N. Tolstikhin, *Chem. Geol.* **145**, 233 (1998).
19. R. Dasgupta, M. M. Hirschmann, N. D. Smith, *Geology* **35**, 135 (2007).
20. R. Dasgupta, M. M. Hirschmann, *Nature* **440**, 659 (2006).
21. D. J. Frost, C. A. McCammon, *Annu. Rev. Earth Planet. Sci.* **36**, 389 (2008).
22. S. S. Shcheka, M. Wiedenbeck, D. J. Frost, H. Keppler, *Earth Planet. Sci. Lett.* **245**, 730 (2006).
23. W. G. Minarik, E. B. Watson, *Earth Planet. Sci. Lett.* **133**, 423 (1995).
24. T. Hammouda, D. Laporte, *Geology* **28**, 283 (2000).
25. T. Kojima, Y. Miyazaki, K. Nomura, K. Tanimoto, *J. Electrochem. Soc.* **155**, F150 (2008).
26. R. J. Sweeney, *Earth Planet. Sci. Lett.* **128**, 259 (1994).
27. C. Simonnet, J. Phalippou, M. Malki, A. Grandjean, *Rev. Sci. Instrum.* **74**, 2805 (2003).
28. Materials and methods are available as supporting material on *Science* Online.
29. D. P. Dobson *et al.*, *Earth Planet. Sci. Lett.* **143**, 207 (1996).
30. M. J. Genge, G. D. Price, A. P. Jones, *Earth Planet. Sci. Lett.* **131**, 225 (1995).
31. Q. Liu, R. A. Lange, *Contrib. Mineral. Petrol.* **146**, 370 (2003).
32. A mantle peridotite with 0.35 volume % of carbonatite melts yields a conductivity of 1 S m^{-1}, which matches the highest value detected by geophysicists at a depth of 150 km (*5*). Such value, however, designates the conductivity in the direction parallel to the ridge axis, whereas the conductivity in the perpendicular direction is lower by a factor of 10 (*5*). We do not incorporate such anisotropic effects in our interpretation, which implies that on average 0.1 volume % of carbonatite melts (300 ± 200 ppmw CO_2) would be needed to explain such mantle conductivities.
33. We thank an INSU grant and the Electrovolc project, which is funded by the French national agency for research (ANR): contract JC05-42707 allocated to F.G. We also thank two anonymous reviewers.

Tight Regulation of Unstructured Proteins: From Transcript Synthesis to Protein Degradation

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

Altered abundance of several intrinsically unstructured proteins (IUPs) has been associated with perturbed cellular signaling that may lead to pathological conditions such as cancer. Therefore, it is important to understand how cells precisely regulate the availability of IUPs. We observed that regulation of transcript clearance, proteolytic degradation, and translational rate contribute to controlling 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 indicate that the availability of IUPs can be finely tuned. Fidelity in signaling may require that most IUPs be 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 to as intrinsically 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) an increased interaction surface area, (ii) conformational flexibility to interact with several targets, (iii) the presence of molecular recognition elements that fold upon binding, (iv) accessible posttranslational modification sites, and (v) the 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*) [supporting online material (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, overexpression of thyroid cancer 1 (TC-1) (*9*) or underexpression of adenosine

[1]Medical Research Council (MRC) Laboratory of Molecular Biology, Hills Road, Cambridge CB2 0QH, UK. [2]Institute for Theoretical Biology, Charité, Humboldt-University, Berlin 10115, Germany. [3]Institute for Biotechnology and Bioengineering Centre for Molecular and Structural Biomedicine, University of Algarve, Faro 8005-139, Portugal.

*To whom correspondence should be addressed. E-mail: madanm@mrc-lmb.cam.ac.uk (M.M.B.); jgsponer@mrc-lmb.cam.ac.uk (J.G.)

5′-diphosphate (ADP) ribosylation factor (Arf) (*10*) and p27 (*11*) has been linked with various types of cancer. Similarly, overexpression of α-synuclein and tau proteins 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 that describe most of the regulatory steps that influence protein synthesis or degradation (table S1 and fig. S1), and we compared the distributions of the values for the proteins in the U and S groups with statistical tests (*14*).

We compared the transcription of genes encoding highly unstructured proteins with 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 ($P = 1 \times 10^{-6}$, Wilcoxon test) (Fig. 2B). The mRNA half-lives of the transcripts that encode highly unstructured proteins were lower than transcripts that encode 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 ($P = 1 \times 10^{-8}$, 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 analyzed polyadenylate [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, which was 68 transcripts (*z* score =

5.3; $P = 5 \times 10^{-10}$) (*14*). Thus, poly(A) tail length and interaction with specific RNA-binding proteins may modulate the 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, and SOM text S2 and S5). The rate of protein synthesis was significantly lower ($P < 10^{-16}$, Wilcoxon test) (Fig. 2E) and protein half-life was shorter ($P = 1 \times 10^{-15}$, Wilcoxon test) (Fig. 2G and SOM text S5) for highly unstructured than for more structured proteins. Two pathways that mediate ubiquitin 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 and figs. S3 and S4), a significantly greater fraction of the unstructured proteins contained PEST motifs (regions rich in proline, glutamic acid, serine, and threonine) ($P < 10^{-16}$, Fisher's exact test) (Fig. 2H and 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 (for example, p27), posttranslational 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 posttranslationally 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 ($P = 1 \times 10^{-12}$, Wilcoxon test) (SOM text S4 and fig. S5). On average, $51 \pm 19\%$ (SD) of all substrates of the kinases are highly unstructured, whereas only $19 \pm 13\%$ (SD) are highly structured [the remaining $30 \pm 14\%$ (SD) of all substrates are

moderately unstructured]. This is a significant bias as compared with the expected genomewide 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 (for example, Cdc28) or activated upon exposure to particular stimuli (for example, Fus3) or stress (for example, Atg1). This suggests that posttranslational 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 ($P = 8 \times 10^{-7}$, Fisher's exact test) (Fig. 2I). At the protein level, analysis of direct experimental data revealed that unstructured proteins have lower noise levels than do structured proteins ($P = 3 \times 10^{-11}$, 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 and tables S3

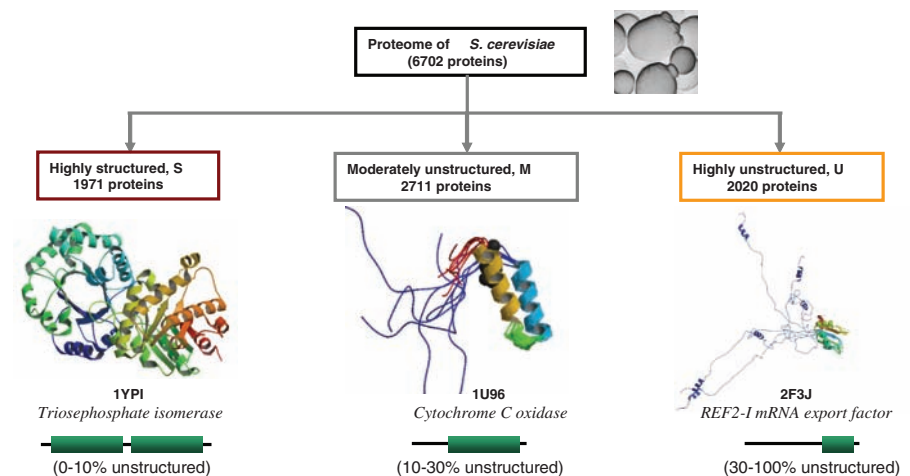## Grouping proteins in the proteome of yeast

Fig. 1. The *S. cerevisiae* proteome was grouped into three categories, highly structured (S), moderately unstructured (M), and highly unstructured (U), based on the fraction of unstructured residues over the entire protein length.

and 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, localization within the major subcellular 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 with two independent statistical tests, the Wilcoxon rank-sum test and the Kolmogorov-Smirnov test (tables S3 to S5) (14). Thus, the reported trends appear to be attributable to the intrinsically unstructured nature of the proteins. Of all the IUPs, those that contain polyglutamine

[poly(Q)] or polyasparagine [poly(N)] stretches seem to be more tightly controlled (table S3 and S4).

Proteins with unstructured regions predominantly have signaling or regulatory roles and are often reused 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 signaling networks leading to dampened or inappropriate activities (19). Spatial and temporal segregation of signaling proteins as well as an increased signaling complexity may contribute to fidelity in regulation (SOM text S5 and fig. S6). In addition, tight regulation of signaling and regulatory IUPs could minimize the potentially harmful effects of ectopic

interactions and may provide robustness to signaling processes by ensuring that such proteins are present in appropriate amounts and time periods. Indeed, free IκB [an IUP that interacts with and inhibits the nuclear factor κB (NF-κB) transcription factor] must be continuously degraded to allow for rapid and robust NF-κB activation and to make the pathway sensitive for a signaling input (20, 21). In contrast, stabilization of free IκB by removal of the PEST motif reduces NF-κB activation (20, 21). In mammalian cells exposed to mild endoplasmic reticulum stress, survival is favored as a direct consequence of the intrinsic instability of mRNAs and of proteins that drive apoptosis (such as Chop, an IUP) as compared with that of proteins that promote adaptation and cell longevity (for example, the chaperone protein BiP, a structured protein) (22).
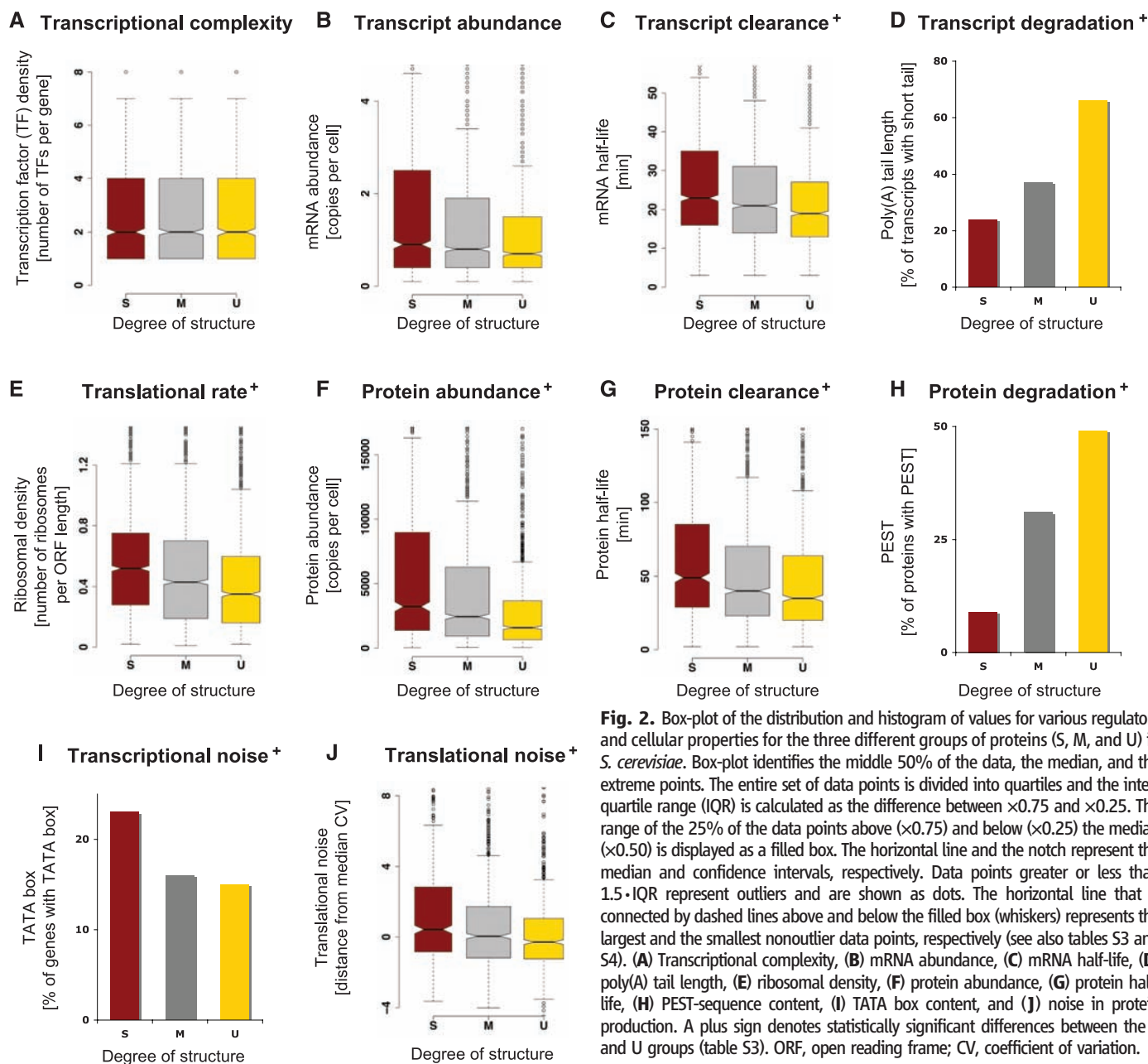

Fig. 2. Box-plot of the distribution and histogram of values for various regulatory and cellular properties for the three different groups of proteins (S, M, and U) in *S. cerevisiae*. Box-plot identifies the middle 50% of the data, the median, and the extreme points. The entire set of data points is divided into quartiles and the inter-quartile range (IQR) is calculated as the difference between ×0.75 and ×0.25. The range of the 25% of the data points above (×0.75) and below (×0.25) the median (×0.50) is displayed as a filled box. The horizontal line and the notch represent the median and confidence intervals, respectively. Data points 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 nonoutlier data points, respectively (see also tables S3 and S4). (**A**) Transcriptional complexity, (**B**) mRNA abundance, (**C**) mRNA half-life, (**D**) poly(A) tail length, (**E**) ribosomal density, (**F**) protein abundance, (**G**) protein half-life, (**H**) PEST-sequence content, (**I**) TATA box content, and (**J**) noise in protein production. A plus sign denotes statistically significant differences between the S and U groups (table S3). ORF, open reading frame; CV, coefficient of variation.
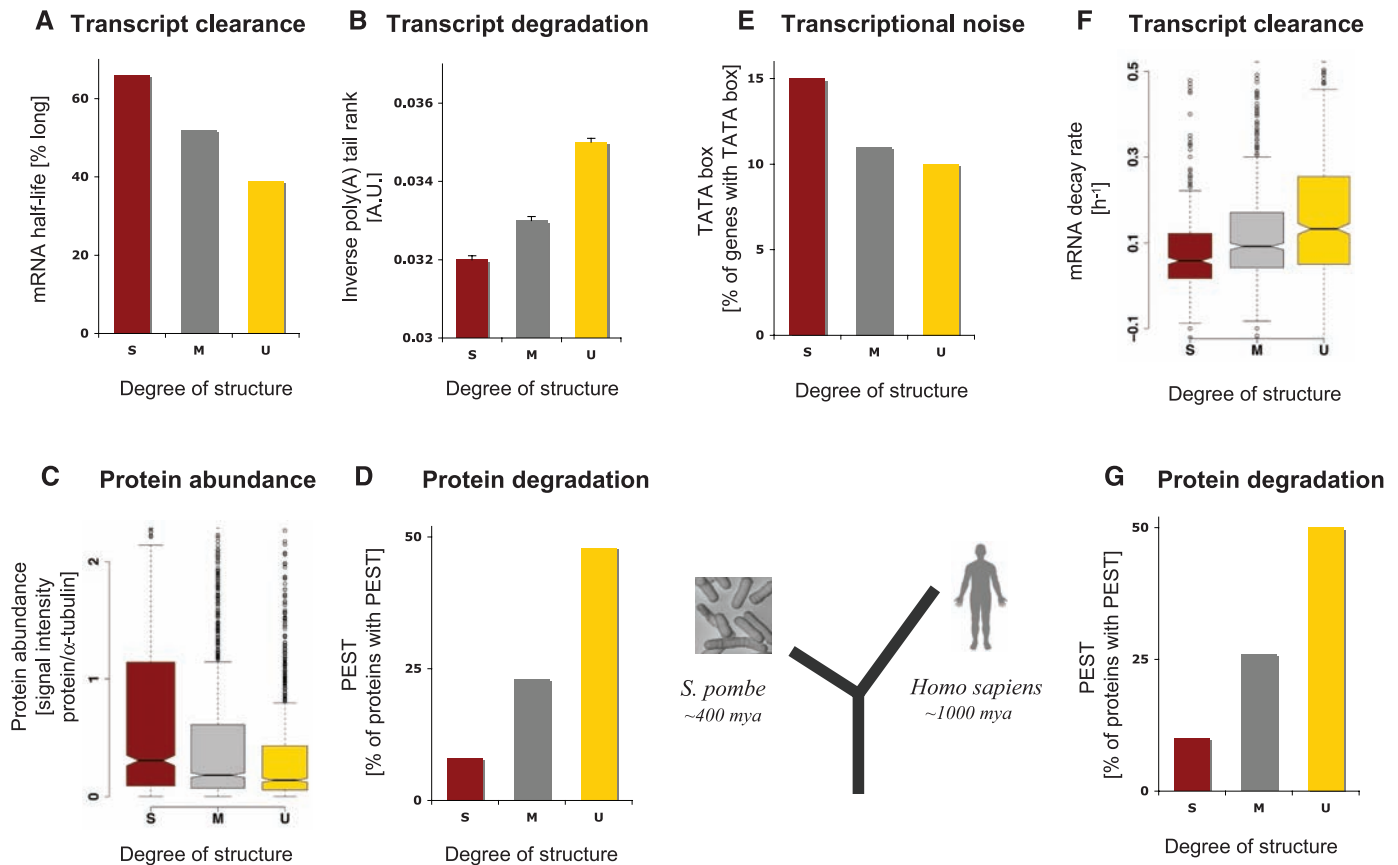
Fig. 3. Box-plot of the distribution and histogram of values for various regulatory and cellular parameters for the three groups of proteins (S, M, and U) in *S. pombe* (**A** to **D**) and humans (**E** to **G**). All reported differences are statistically significant (table S4). A.U., arbitrary units; mya, million years ago.

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 lifetime of a cell. Fine-tuning of IUP availability can be achieved with posttranslational 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 (for example, 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 20*S* 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 posttranslational 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 that describe 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 may 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*).

References and Notes
1. A. K. Dunker *et al.*, *J. Mol. Graph. Model.* **19**, 26 (2001).
2. P. E. Wright, H. J. Dyson, *J. Mol. Biol.* **293**, 321 (1999).
3. R. W. Kriwacki, L. Hengst, L. Tennant, S. I. Reed, P. E. Wright, *Proc. Natl. Acad. Sci. U.S.A.* **93**, 11504 (1996).
4. P. Tompa, *FEBS Lett.* **579**, 3346 (2005).
5. C. J. Oldfield *et al.*, *Biochemistry* **44**, 12454 (2005).
6. J. J. Ward, J. S. Sodhi, L. J. McGuffin, B. F. Buxton, D. T. Jones, *J. Mol. Biol.* **337**, 635 (2004).
7. J. Liu *et al.*, *Biochemistry* **45**, 6873 (2006).
8. L. M. Iakoucheva, C. J. Brown, J. D. Lawson, Z. Obradovic, A. K. Dunker, *J. Mol. Biol.* **323**, 573 (2002).
9. M. Sunde *et al.*, *Cancer Res.* **64**, 2766 (2004).
10. C. J. Sherr, *Nat. Rev. Cancer* **6**, 663 (2006).
11. M. Grimmler *et al.*, *Cell* **128**, 269 (2007).
12. F. Chiti, C. M. Dobson, *Annu. Rev. Biochem.* **75**, 333 (2006).
13. M. Goedert, *Nat. Rev. Neurosci.* **2**, 492 (2001).
14. Materials and methods are available as supporting material on *Science* Online.
15. P. Tompa, J. Prilusky, I. Silman, J. L. Sussman, *Proteins* **71**, 903 (2007).
16. I. Chu *et al.*, *Cell* **128**, 281 (2007).
17. L. M. Iakoucheva *et al.*, *Nucleic Acids Res.* **32**, 1037 (2004).
18. J. M. Raser, E. K. O'Shea, *Science* **304**, 1811 (2004).
19. T. Pawson, N. Warner, *Oncogene* **26**, 1268 (2007).
20. E. Mathes, E. L. O'Dea, A. Hoffmann, G. Ghosh, *EMBO J.* **27**, 1421 (2008).
21. E. L. O'Dea *et al.*, *Mol. Syst. Biol.* **3**, 111 (2007).
22. D. T. Rutkowski *et al.*, *PLoS Biol.* **4**, e374 (2006).
23. H. K. Chen *et al.*, *Cell* **113**, 457 (2003).
24. I. M. Chu, L. Hengst, J. M. Slingerland, *Nat. Rev. Cancer* **8**, 253 (2008).
25. G. Asher, N. Reuven, Y. Shaul, *Bioessays* **28**, 844 (2006).
26. X. Li *et al.*, *Mol. Cell* **26**, 831 (2007).
27. W. S. Tambyrajah, L. D. Bowler, C. Medina-Palazon, A. J. Sinclair, *Arch. Biochem. Biophys.* **466**, 186 (2007).
28. G. K. Tofaris, R. Layfield, M. G. Spillantini, *FEBS Lett.* **509**, 22 (2001).
29. D. C. David *et al.*, *J. Neurochem.* **83**, 176 (2002).
30. W. E. Balch, R. I. Morimoto, A. Dillin, J. W. Kelly, *Science* **319**, 916 (2008).
31. The authors acknowledge MRC for funding and 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 apologize for, because of lack of space, not citing the work that describes the datasets and other references. Many can be found in the SOM. M.M.B. acknowledges Darwin College and Schlumberger Ltd, and M.M.B. and M.E.F. acknowledge the Royal Society for support. J.G. is funded by an MRC Special Training Fellowship in computational biology.