7666–7675 Nucleic Acids Research, 2012, Vol. 40, No. 16 Published online 20 June 2012
doi:10.1093/nar/gks511

eIF4E-binding protein regulation of mRNAs with differential 5′-UTR secondary structure: a polyelectrostatic model for a component of protein–mRNA interactions

Andrew Cawley and Jim Warwicker*

Faculty of Life Sciences, University of Manchester, Manchester Interdisciplinary Biocentre, 131 Princess Street, Manchester M1 7DN, UK

Received September 22, 2011; Revised April 20, 2012; Accepted May 8, 2012

ABSTRACT

Control of translation in eukaryotes is complex, depending on the binding of various factors to mRNAs. Available data for subsets of mRNAs that are translationally up- and down-regulated in yeast eIF4E-binding protein (4E-BP) deletion mutants are coupled with reported mRNA secondary structure measurements to investigate whether 5′-UTR secondary structure varies between the subsets. Genes with up-regulated translational efficiencies in the *caf20Δ* mutant have relatively high averaged 5′-UTR secondary structure. There is no apparent wide-scale correlation of RNA-binding protein preferences with the increased 5′-UTR secondary structure, leading us to speculate that the secondary structure itself may play a role in differential partitioning of mRNAs between eIF4E/4E-BP repression and eIF4E/eIF4G translation initiation. Both Caf20p and Eap1p contain stretches of positive charge in regions of predicted disorder. Such regions are also present in eIF4G and have been reported to associate with mRNA binding. The pattern of these segments, around the canonical eIF4E-binding motif, varies between each 4E-BP and eIF4G. Analysis of gene ontology shows that yeast proteins containing predicted disordered segments, with positive charge runs, are enriched for nucleic acid binding. We propose that the 4E-BPs act, in part, as differential, flexible, polyelectrostatic scaffolds for mRNAs.

INTRODUCTION

The initiation of eukaryotic translation is subject to multiple mechanisms for regulation (1,2). Binding of a 43S pre-initiation complex to the capped 5′-end of mRNA is facilitated by the cap-binding factor eIF4E and partners eIF4G and eIF4A in the eIF4F complex. The helicase subunit of eIF4F, eIF4A, promotes scanning of mRNAs containing secondary structure in their 5′-UTRs (3). Poly(A) binding protein (PABP) interacts also with eIF4G and can mediate the formation of a circular messenger ribonucleoprotein (mRNP) complex by linking the cap and the poly(A) tail. Because the rate of translation initiation depends on the recognition of the mRNA 5′-cap by eIF4F, this process is central for control of translation (1). 4E-binding proteins (4E-BPs) compete with eIF4G for a common binding site on the companion eIF4F subunit, eIF4E (4), and are therefore inhibitory for translation. Regulatory mechanisms generally fall into two categories, those impacting on the eIFs (e.g. by protein phosphorylation) or ribosomes, which affect initiation overall, and those acting on mRNAs, with the potential to be mRNA subset-specific (2). In the latter category, interaction of specific RNA-binding proteins is more common for the 3′-UTR than the 5′-UTR and follows a general mechanism in which the specifically (3′-UTR) bound protein cross-links through an intermediate protein to a cap-binding protein, forming an inhibitory loop that precludes access for eIF4F (2).

Genomic studies are facilitating analysis of translational control mechanisms, with yeast a key model organism. For example, eIF4G depletion in yeast narrows the range of translational efficiencies, rather than preventing translation, for most genes (5). Translation efficiencies with the greatest dependence on eIF4G were found for mRNAs that are relatively well translated, without long or highly structured 5′-UTRs. This result is consistent with eIF4F being more important in enhancing 43S attachment to the mRNA 5′-end than in scanning through long, structured 5′-UTRs (5). Another study has looked at the effects on translation of deletion of the two eIF4G isoforms in yeast (6). While there is no clear functional

*To whom correspondence should be addressed. Tel: +44 161 306 4490; Fax: +44 161 275 5082; Email: jim.warwicker@manchester.ac.uk

differentiation between the eIF4G isoforms, mRNAs with longer poly(A) tails are more sensitive to eIF4G loss, consistent with a coupling between translation efficiency and eIF4G–PABP interaction (6).

In addition to modulating translation through binding proteins (2,7), the intrinsic properties of flanking UTRs may also influence translation (8). For example, analysis of UTRs in sets of yeast genes up- and down-regulated at translation in response to stress shows that the 5′-UTRs of up-regulated mRNAs are relatively longer, with an over-representation of upstream open reading frames (ORFs) (9). The secondary structure of mRNA has been predicted at the translational start site. An early study of eukaryotic and prokaryotic mRNAs found relatively low predicted secondary structure at translation initiation sites (10), an observation repeated with a more extensive analysis of yeast mRNAs (11). A broad study of predicted mRNA secondary structure in 340 species reported a similar reduction in secondary structure near the start codon (12). There have also been reports that 5′-UTR structures influence translation. Genes with relatively long and structured 5′-UTRs are translated more slowly (13) and, conversely, 5′-UTRs predicted to be weakly folded lead to higher translation rates (14).

Understanding of the structural biology and biophysics of eukaryotic translation has advanced with the determination of atomic resolution structures for several domains of initiation factors and regulators (15,16) and from an improved understanding of structure–function in the 43S pre-initiation complex (17). Intrinsically disordered protein regions play a role in cap-dependent translation. Parts of eIF4E undergo folding transitions upon cap binding and eIF4G binding (18). In eIF4G, beyond folding transitions centred on a specific interface with eIF4E, there exist RS (arginine–serine)-rich domains that interact with mRNA (19), which have been implicated in promoting assembly of eIF4F–mRNA complexes (20). RS domains are widely involved in protein–RNA interactions (21) and are believed to be intrinsically disordered (22). In previous work, we have simulated non-specific interactions between the charged surfaces of eIF4E/eIF4G structured domains and a polymer bead model for mRNA, suggesting that these interactions have the potential to supplement the cap–eIF4E interaction (23) and that mRNA secondary structure could, in principle, influence the interaction with protein charges.

The aim of this study is to investigate the potential for weak protein–mRNA interactions making use of genomic data. Translational profiling has revealed that two yeast 4E-BPs, Eap1p and Caf20p, modulate the translation of more than 1000 genes, with evidence that mRNA-binding proteins are in part responsible, through complexation with one or other 4E-BP (24). Genome-wide measurement of RNA secondary structure in yeast has also been reported, in which the parallel analysis of RNA secondary structure (PARS) technique is based on structure-specific nuclease treatment and subsequent sequencing (25). Combining these studies, it is possible to determine whether there are secondary structural differences between the mRNA subsets whose translation is modulated by the 4E-BPs. The structural framework in which to interpret these results is limited, because mRNA is a flexible polymer and large segments of the 4E-BPs have the properties of intrinsically disordered domains. In place of structural models, we complement bioinformatics analysis of mRNA secondary structures with a parallel study of charged amino acid runs in 4E-BPs, eIF4G and in the context of the yeast proteome. Relatively high 5′-UTR secondary structure is found for the translationally up-regulated set of mRNAs associated with one of the 4E-BP deletions. On the basis of a known role for eIF4G positive charge runs in mRNA binding and a gene ontology analysis for the yeast proteome, it is suggested that positively charged regions in the 4E-BPs could be associated with mRNA binding.

MATERIALS AND METHODS

Yeast gene data sets and mRNA secondary structures

Sets of genes up- and down-regulated at translation, measured by polyribosome association, upon deletion of either of the two yeast (*Saccharomyces cerevisiae*) 4E-BPs, Caf20p and Eap1p, were derived from previous work (24). Table 1 gives the numbers of genes in these four subsets, before (N_{red}) and after ($N_{non\text{-}red}$) removal of those (redundant) genes that appear in more than one subset and removal of entries from duplicate probesets (24). Secondary structures for *S. cerevisiae* mRNAs, measured with the PARS method (25), were obtained from http://genie.weizmann.ac.il/pubs/PARS10/pars10_catalogs.html. Untranslated regions were mapped onto the mRNAs

Table 1. UTR lengths and PARS averages for non-redundant gene subsets

	N_{red}	$N_{non\text{-}red}$	$N_{5'\text{-}UTR}$	$N_{3'\text{-}UTR}$	5′-UTR len	3′-UTR len	5′-UTR PARS	3′-UTR PARS
All			2680	2883	78	126	0.059	0.026
caf20Δ up	490	408	190	201	**117 (3.3e-8)**	158 (0.230)	**0.202 (5.2e-7)**	0.057 (0.901)
caf20Δ down	374	305	87	94	65 (0.801)	**117 (4.6e-4)**	0.050 (0.258)	**−0.027 (0.033)**
eap1Δ up	231	154	62	68	77 (0.687)	**124 (1.6e-3)**	0.062 (0.924)	0.054 (0.431)
eap1Δ down	277	204	128	140	82 (0.092)	139 (0.529)	0.154 (0.075)	0.016 (0.184)

'All' denotes all yeast genes in the PARS data (25) that have defined 5′-UTR or 3′-UTR coordinates; N_{red}, $N_{non\text{-}red}$ are the numbers of genes up- and down-regulated in the 4E-BP deletion mutants before and after removing those genes appearing in more than one subset, as well a small number of entries from duplicate probesets (24); 5′,3′-UTR len and PARS columns give length and PARS score averages over genes in the given gene set, along with (in brackets) the *P*-value associated for a Mann–Whitney test of similarity between the up-/down-regulated (non-redundant) subset and the 'All' set. Pairings that are significantly different at the 5% level are in bold.

using coordinates (26) available from the same location. The PARS technique profiles secondary structure of mRNAs by deep sequencing of fragments after treatment with structure-specific enzymes. RNase V1 preferentially cleaves double-stranded RNA, and S1 nuclease preferentially cleaves single-stranded RNA. For each nucleotide, the logarithm of the ratio between the number of reads obtained for that nucleotide in the V1-treated sample and that obtained in the S1-treated sample is computed. The PARS score of a particular nucleotide is then defined as this logarithm, for reads with the first base observed as the nucleotide immediately downstream of the site in question (25). The ability of PARS methodology to detect base pairing and secondary structure formation has been tested in relation to RNA footprinting data and known structures, with significant correspondence between PARS and computational predictions (25).

PARS processing for 5′-UTRs and 3′-UTRs

Code was written in Perl for various stages of data processing. Subsets of genes with up- or down-regulation, associated with 4E-BP deletions (24), are provided as a text file to make_nonredundant.pl to produce non-redundant versions (i.e. each gene is associated with just one subset). Either the redundant or non-redundant data are then used in code, PARS-profiles.pl, which associates each gene with UTR coordinate data and PARS scores (25), where available. Profiles are produced traversing the 5′- and 3′-UTRs either from the mRNA termini or from the start/stop codons. Several features of this profiling can be adjusted, including the number of nucleotides sampled (from mRNA termini and start/stop) and whether the profiles are smoothed. Smoothing is achieved, at each nucleotide location, by averaging the PARS scores over a window of 11 nt centred on the selected nucleotide. The window is reduced at the UTR termini. Smoothing is applied unless stated otherwise. We generally study the profiles using the mRNA termini as origin. A profile for all genes, aligned to an origin at the start codon, matches with the original report [Figure 3c (25), data not shown]. There is an option to exclude a given number (typically 20) of nucleotides adjacent to the start codon, in the 5′-UTR, and adjacent to the stop codon, in the 3′-UTR. This is designed to prevent systematic variations in the profiles averaged within a subset, where UTRs will contribute lower PARS at the start/stop codon (25). These contributions will be out of phase, owing to variation in UTR lengths, when the profile is calculated traversing from either of the mRNA termini. PARS-profiles.pl produces profiles that are averaged over all genes in a subset (that can be mapped with UTR coordinates and which have PARS scores), and over all genes (with UTR coordinates and PARS scores). Separate code (write_PARS_individual.pl) allows output of individual UTR PARS profiles (i.e. not averaged over genes). For statistical comparison of a subset PARS profile and the overall profile, loop_resample.pl selects random subsets from the overall gene/UTR data, with sample number matching the subset size. The subset profile is then compared to the random sampling.

RNA-binding protein motifs

Past work has identified nucleotide sequence motifs that are enriched in mRNAs bound by specific RNA-binding proteins (RBPs) in yeast (27,28). RNA motifs, with the most significant enrichments, for 14 RBPs have been derived (27). The 5′-UTRs of mRNAs that are translationally up-regulated in the *caf20Δ* mutant were searched for occurrence of these 14 motifs. For each motif match, a PARS score average was calculated across the nucleotides of the motif. These PARS averages were then assigned to a single 10 nt bin, starting from the 5′-end of the 5′-UTR, according to the location of the central nucleotide within a motif. If the motif contained an even number of nucleotides, the closest nucleotide below centre was used for binning. The sum of the PARS averages assigned to a bin were divided by the number of motif matches in that bin and the resultant histogram plotted. In addition, the number of motif matches in each bin was plotted. The total number of motif matches, for the 14 RBPs, in the 5′-UTRs of genes translationally up-regulated in the *caf20Δ* mutant, is 80.

Charge runs in yeast proteins

Structural disorder for amino acid sequences was predicted with the GlobPlot set of amino acid propensities (29), coded locally, with a window of 21 amino acids giving an average value for the central amino acid in that window. Regions of disorder computed in this way were comparable to those computed with the Fold Index scheme (30). Perl code (seq_props_all.pl) that calculates disorder over 21 amino acid windows also gives the net charge in each window. It is assumed that the relevant cytoplasmic environment is close to neutral pH, so that the sidechains of aspartic and glutamic acids carry −1e charge, and lysine and arginine +1e. Histidine (intrinsic pKa of 6.3) is likely to be neutral in the absence of favourable interactions elevating the pKa. A typical value for pKa deviation upon salt-bridge formation is 1 pH unit (31). Histidine is treated as either neutral in the charge run calculations or with the effect of elevated histidine pKa (e.g. through interaction with the polyphosphate backbone of a nucleic acid). Charge runs were calculated for a set of 5885 proteins, the translations of all systematically named ORFs, obtained from the *Saccharomyces* Genome Database.

Caf20p, Eap1p, Tif4631p and Tif4632p amino acid sequences were searched against the Protein Data Bank (PDB) (32) with BLAST (33) to identify regions with known 3D structural domains. Two such regions were identified for eIF4G, with PDB ids 2vso (eiF4G–eIF4A complex) and 1rf8 (eIF4G–eIF4E) selected to display these regions and their interactions. Amino acid conservation was studied with the ConSurf server (34) (http://consurf.tau.ac.il/).

Gene ontology

Sets of yeast genes with charge runs of a certain net charge in a 21 amino acid window were analyzed for enrichment of molecular function gene ontology (GO) terms with the

GO-TermFinder software (35) implemented at Princeton (http://go.princeton.edu/cgi-bin/GOTermFinder). The set of 5885 ORFs used in the charge run analysis was provided as background for the enrichment calculation. Parameters were otherwise set at default values, including use of the Bonferroni correction for multiple comparisons. Enrichment for a particular GO molecular function (e.g. nucleic acid binding) was calculated as the ratio of the percentage of the test set genes associated with the molecular function to the percentage of the background set associated with the same molecular function.

RESULTS AND DISCUSSION

PARS scores of UTRs for mRNA subsets translationally up-/down-regulated in 4E-BP gene deletions

In addition to the numbers of genes translationally up- and down-regulated, upon 4E-BP gene deletion, (before and after redundancy imposition), Table 1 gives the average lengths and PARS scores for 5′-UTRs and 3′-UTRs in these subsets. Statistics for subset differences to the overall yeast gene set (that match to PARS scores and UTR coordinates) are shown. The 5′-UTR properties, length and PARS score, for genes up-regulated in the *caf20Δ* mutant, show the most difference to the overall set. Here, the average 5′-UTR length is longer and the average 5′-UTR PARS score higher. Table 2 shows the equivalent UTR length and PARS score data to those in Table 1 but for redundant subsets. A similar pattern of results is obtained with 5′-UTR length and PARS score for the *caf20Δ* mutant up-regulated subset again giving the most significant deviations from the overall set.

PARS score profiles averaged over UTR subsets for 4E-BP gene deletions

Following the results in Table 1, highlighting 5′-UTRs, averaged PARS profiles are shown in Figure 1. The profiles are displayed with origins fixed at the 5′ terminus of each mRNA and with 20 UTR nucleotides adjacent to the start codon excluded from the analysis. Each plot gives the PARS profile for one deletion mutant, compared with the analogous profile for all genes, and with a spread of standard deviations for 1000 random resamples corresponding to the subset number of UTRs. In agreement with Table 1, the 5′-UTRs of genes translationally up-regulated in the *caf20Δ* mutant show higher PARS scores (more secondary structure) on average.

There are other differences between subset and overall PARS profiles in Figure 1. The accuracy of these comparisons reduces with distance from the 5′ end, as the number of UTRs included in the calculations decreases. To examine the influence of UTR length on profile differences, Figure 2 compares 5′-UTR results for the *caf20Δ* mutant [panel (a), as in Figure 1], with nucleotide exclusion from the start codon increased from 20 to 40 [panel (b)]. In addition, the entire set of genes was subjected to a pre-screen for 5′-UTRs of length $\geq$100, as well as implementing the 40-nt exclusion from start codon (Figure 2c). The average PARS score for the subset profile is not reduced by these measures, suggesting that increased secondary structure in the *caf20Δ* mutant set is not the result of differential sampling of 5′-UTR secondary structure properties immediately adjacent to the start codon. It may be that the significant increases in 5′-UTR length and average PARS scores, for genes up-regulated in the *caf20Δ* mutant, are related. For example, longer 5′-UTRs may simply give more scope for the development of mRNA secondary structure. It is not apparent that GC content lies behind the PARS profile differences because it is relatively uniform for subset averages over 5′-UTRs at 33–35%, comparable with previous bioinformatics analysis (36).

Comparison of sequences for yeast 4E-BPs and eIF4G

To investigate whether there are protein sequence features that could aid understanding of PARS score profile data, the sequences of Caf20p, Eap1p and Tif4631p were studied. The second eIF4G, Tif4632p, has similar properties to Tif4631p and is not shown. Amino acid sequences are aligned at the eIF4E-binding motif, YxxxxLφ, and disorder and net charge (with neutral histidine sidechains) shown for windowed calculations over 21 amino acids (Figure 3). Two regions of eIF4G are structurally annotated. Representative structures are displayed (as solid surfaces in complexes with other eIF4F components). These regions map, as expected, to low predicted structural disorder. Additionally, the sequence is more negatively charged (red) for each region, consistent with the surface plots. Three eIF4G regions with non-specific mRNA binding affinity (19) are marked in Figure 3 as RNA1, RNA2, RNA3 (20) along with the PABP-binding region. The non-specific mRNA-binding sites correspond to positive charge runs. RNA2 and RNA3 were identified as RS sites (arginine-serine rich) (19), and RNA1 is also clearly positively charged. All three RNA-binding regions are predicted to be

Table 2. UTR lengths and PARS averages for redundant gene subsets

	5′-UTR len	3′-UTR len	5′-UTR PARS	3′-UTR PARS
All	78	116	0.059	0.026
caf20Δ up	**114 (2.4e-9)**	**158 (0.038)**	**0.188 (7.6e-7)**	0.066 (0.196)
caf20Δ down	66 (0.488)	**113 (3.8e-5)**	0.057 (0.158)	−0.016 (0.090)
eap1Δ up	81 (0.807)	**131 (1.6e-3)**	0.075 (0.904)	0.068 (0.141)
eap1Δ down	84 (0.060)	136 (0.934)	0.148 (0.071)	0.022 (0.294)

Columns as described for Table 1 but with data for redundant gene subsets.

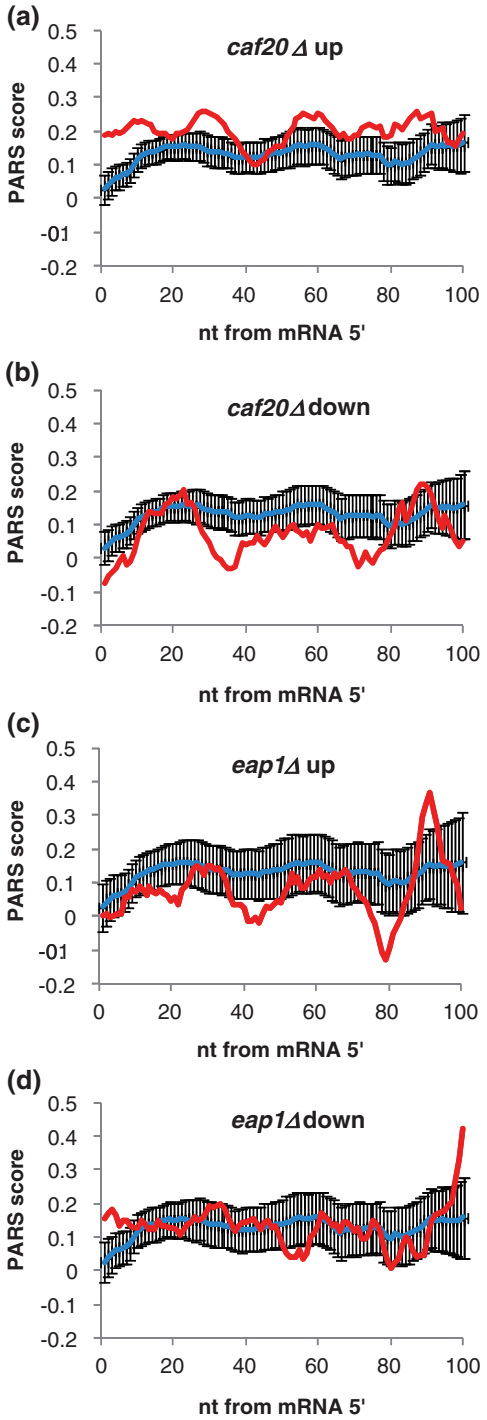

Figure 1. Averaged PARS score profiles for 5′-UTRs. Profiles (red) are averaged over all 5′-UTRs in a subset of up/down translational regulation in 4E-BP deletion mutants. In addition, an 11 nt sliding window is used to average within a 5′-UTR, and 20 nts adjacent to the start codon are excluded from the analysis. The profile averaged over all mRNAs with PARS scores and known 5′-UTR lengths is shown (blue) in all panels. Results of resampling (1000 trials), extracting the same number of mRNAs present in a subset, from all mRNAs, are shown as standard deviation around the overall profile. Profiles are shown for 5′-UTRs in the following translationally regulated mRNA sets: (**a**) *caf20Δ* up-regulated, (**b**) *caf20Δ* down-regulated, (**c**) *eap1Δ* up-regulated and (**d**) *eap1Δ* down-regulated.

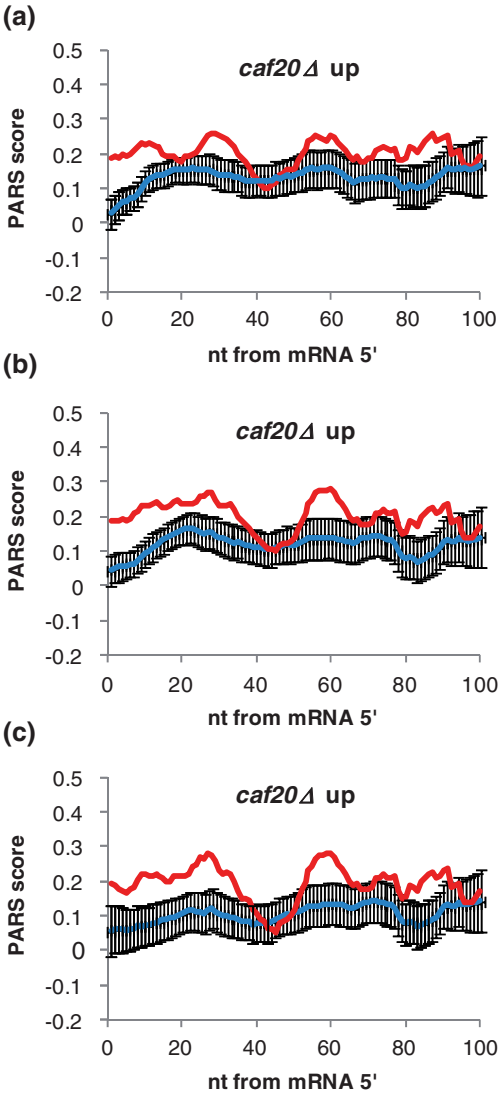

Figure 2. Effects of UTR length on averaged PARS score profiles. Data for 5′-UTRs of mRNAs translationally up-regulated in the *caf20Δ* mutant are calculated as in Figure 1, but with the following variations. (**a**) No variation, same parameters and plot, with origin at the mRNA 5′ end, as Figure 1a. (**b**) The 20-nt exclusion 5′ to the start codon is increased to 40 nts. (**c**) In addition to the 40-nt exclusion, only mRNAs pre-filtered for 5′-UTR length 100 nts or greater are included.

structurally disordered. Furthermore, positive charge is involved in mRNA binding, since arginine replacement with alanine in RNA2 abolishes its RNA-binding activity (19). Deleting these regions impacts cumulatively on eIF4G function, with eIF4G lacking one of RNA1, RNA2 or RNA3 still active, but removal of two or three of these sites giving strong impairment or inactivity (19). This background encourages us to look at potential links between positive charge runs and mRNA binding in the current work.

The RNA2 region of eIF4G is in a similar location (about 100 amino acids C-terminal to the eIF4E-binding motif, YxxxxLφ) to positively charged runs in regions of predicted disorder for Caf20p and Eap1p. There are additional positive charge runs in regions of predicted

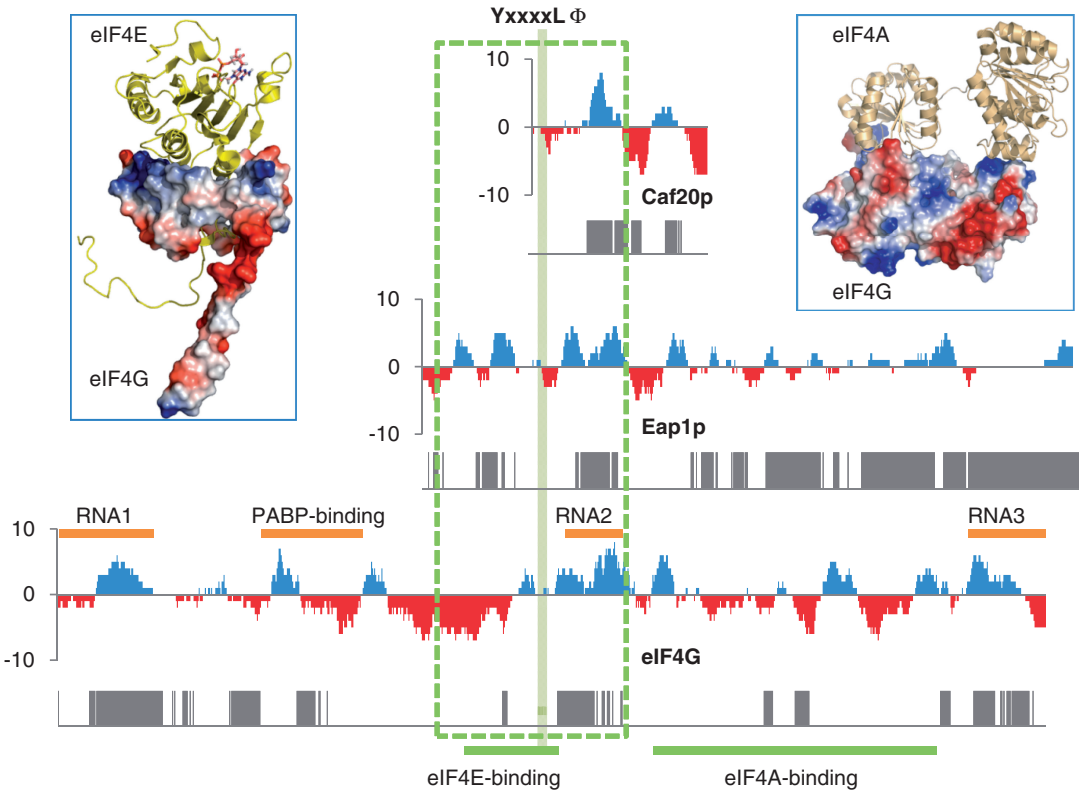

Figure 3. Properties of 4E-BPs and eIF4G. Plots of (windowed) predicted structural disorder and net charge, are shown for each of Caf20p, Eap1p, and eIF4G (Tif4631p). These proteins are scaled to maintain their relative sequence lengths, and aligned on their 4E-binding motifs, marked with the consensus sequence and with a vertical green bar. Charge colour-coding is blue/positive and red/negative for both the sequence plots and molecular surfaces. Predicted disorder is shown by grey vertical lines. Around the 4E-binding motif and therefore proximal to the 5′-UTR of an eIF4E-bound mRNA, a dashed green box marks a region of about 200 amino acids, within which the disorder and net charge properties differ between these 3 proteins. Structural annotation for two eIF4G regions is displayed, in complex with other eIF4F components in each case, and with the eIF4G element drawn as an electrostatic potential coded surface. Both of the structural annotations correspond to eIF4G regions largely devoid of grey blocks, consistent with structured protein. The full 1rf8 coordinates (46) are shown for eIF4G in complex with eIF4E, although the structure of the isolated termini must be uncertain. Locations of the 3D structures in the eI4FG sequence are denoted by green bars. Orange bars give the sites of the PABP-binding region and three RNA-binding segments (19,20) in eIF4G.

disorder for these two 4E-BPs, notably in the much longer Eap1p sequence. The boxed region of about 100 amino acids either side of the specific eIF4E-binding motif (Figure 3) shows differences between the 4E-BPs and eIF4G. Caf20p has little sequence N-terminal to the binding motif, and the dominant C-terminal feature is the positively charged region. Eap1p is predicted to be largely positively charged and disordered on each side, and eIF4G (Tif4631p) is ordered and negatively charged N-terminal, and positively charged and predicted to be disordered C-terminal to the motif. We currently lack structural or simulation models for disordered protein interactions with mRNA, including any potential affect of mRNA secondary structure. It is therefore difficult to draw precise links between the protein sequence properties of Figure 3 and the PARS score profiles (Figures 1 and 2). It is, however, interesting that the most positively charged window in these sequences is a net charge of +8 in Caf20p. This region is relatively well conserved in the context of these protein sequences, particularly the positively charged amino acids (not shown). In addition, this window in Caf20p also contains four histidines that are assigned zero charge in the analysis of Figure 3, but could ionize, increasing the positive charge. Charge profiles with histidine included are studied in a subsequent section.

GO term enrichment for yeast proteins with regions rich in positive charge

To further investigate charge runs in yeast proteins and biological function, an analysis of GO term (molecular function) enrichment in protein-coding ORFs of varying charge properties was carried out. Figure 4a shows the fraction of protein-coding genes that contain 21 amino acid windows (predicted to be ordered or disordered) of the given net charge, positive or negative. Proteins with high negative charge windows are more common than those with positive charge. Taking a net charge of 8 (with neutral histidines), the highest for the proteins in Figure 3 (in Caf20p), the most significant enrichments for both positive and negative charge runs, in the function GO category, are nucleic acid binding. The enrichments are higher for positive charge runs than their negative counterparts. For example, at a net charge of +8, enrichment ratios for nucleic acid binding, DNA binding, RNA binding are 2.62, 2.32 and 1.78, respectively, in windows of predicted disorder. The equivalent enrichment

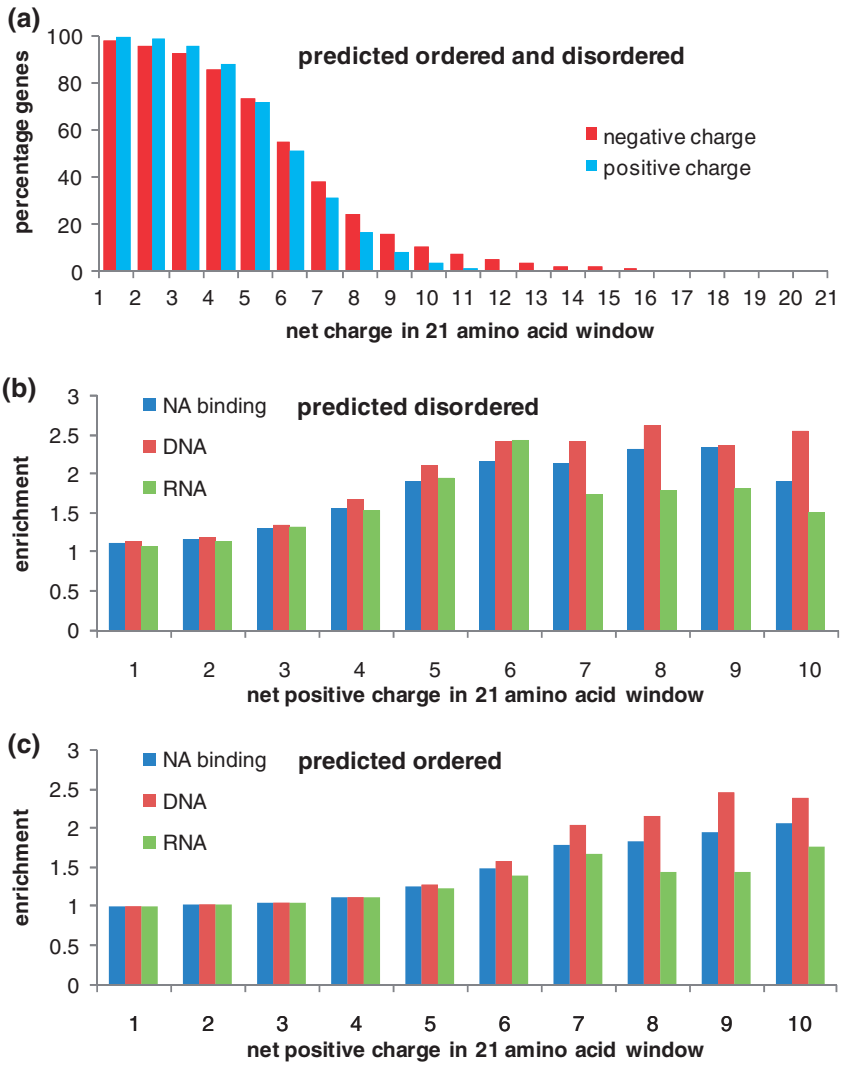

Figure 4. Charge runs in yeast proteins. (**a**) The percentages of protein coding genes containing at least one 21 amino acid window bearing the relevant net charge, positive or negative, are shown. (**b**) and (**c**) Enrichment for the GO function terms nucleic acid binding, DNA binding, RNA binding, in protein subsets that contain at least one 21 amino acid window bearing the listed net positive charge. (b) Regions of predicted disorder; $P < 0.01$ (Bonferroni correction applied) for all enrichments from net charge +2 to +7, inclusive. (c) Regions of predicted order; $P < 0.01$ for all enrichments from +3 to +10, inclusive.

ratios for a net charge of −8 are 1.77, 1.46 and 1.10. Figure 4 shows the enrichment of annotation for nucleic acid binding, and within that RNA- and DNA-binding, for windows with net charge +1 to +10, and for predicted disordered (Figure 4b) and ordered (Figure 4c) segments. Both plots show significant enrichment for nucleic acid binding as net positive charge increases, with this more evident at lower charge for disordered, relative to ordered. These data indicate that windows of net positive charge comparable to those in the 4E-BPs are consistent with nucleic acid binding. A comparison of Figure 4b and c shows that enrichment for RNA-binding proteins falls off more quickly, as net charge increases at higher values, than does that for DNA-binding proteins. This difference is more evident for proteins containing windows predicted to be disordered. Analysis of those genes lost when increasing net charge from +6 to +7 reveals a large number of retrotransposon Ty elements. These contain nucleocapsid proteins, structural components of the virus-like particle shell that encapsulates the retrotransposon RNA (37), involved in non-specific RNA interactions and packaging (38).

Sequence comparisons of 4E-BPs and eIF4G with charged histidine

Having noted that the window of predicted disorder with the highest positive charge in the two 4E-BPs and eIF4G, is also rich in histidines, the charge runs analysis was repeated with ionized histidine. The effect (Figure 5) is to substantially enhance the positive charge profile of the region previously identified in Caf20p, such that it now stands out in magnitude from all other regions in these three proteins. A BLAST search (33) reveals 12 homologues of Caf20p with a sequence identity of ≥55%. There is then a gap to further homologues with sequence identity of ≤33%. The left-hand insert of Figure 5 shows charge profiles for the Caf20p homologues

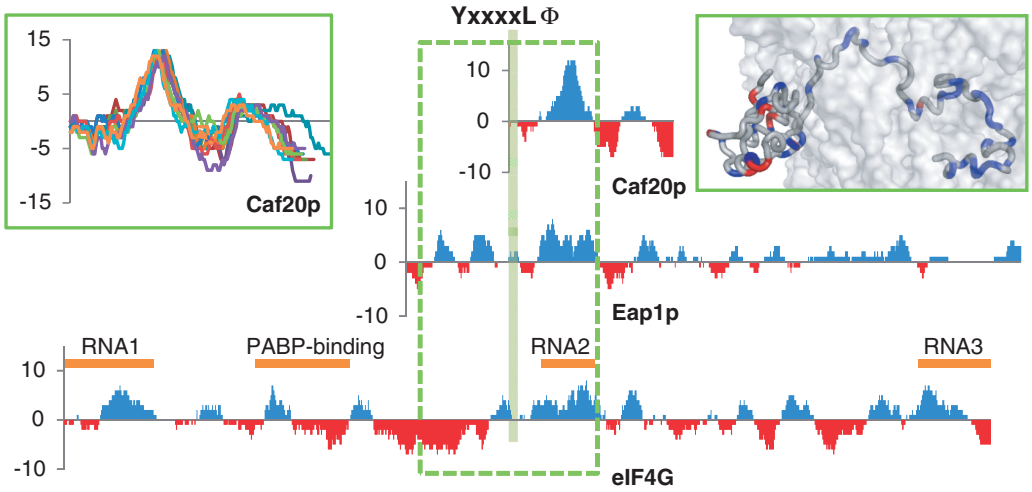

Figure 5. Properties of 4E-BPs and eIF4G, with histidine charge. The charge profiles for Caf20p, Eap1p and eIF4G (Tif4631p) are drawn as in Figure 3, but with histidine sidechains bearing +1 charge. The three RNA binding segments and the PABP binding region are indicated for eIF4G, as well as the location of the eIF4E-binding motif in all three proteins. The left-hand inset shows the charge profiles for Caf20p homologues (see text). In the right-hand inset, yeast ribosomal protein L28 is drawn as a tube (blue basic, red acidic), with a transparent grey surface representing ribosomal RNA [PDB ids 3u5d, 3u5e (39)]. The amino-terminal extension has similar charge and predicted disorder properties to the peak positive region of Caf20p.

of ≥55% sequence identity, aligned at the amino termini. The positively charged region is a relatively well-conserved feature within this set of proteins.

The maximum positive charge of a 21 amino acid window in Figure 5 (with ionized histidines) is +12 in Caf20p. We found 16 other yeast proteins with a maximum charge of +12 or greater, including ionized histidines, in a region of predicted disorder. Searching these against the PDB revealed one structure, for ribosomal protein L28. The positively charged region, that is predicted to be disordered, is an N-terminal extension from a folded L28 domain. It is ordered by virtue of extensive interactions with ribosomal RNA (39). The right-hand inset of Figure 5 shows this L28 extension, with basic amino acids highlighted on a backbone tube. Peak charge for a window in L28 is +12, the same as that for Caf20p.

Caf20p is distinct from Eap1p and eIF4G, in terms of charge profiles, when charged histidine is included. A region with the same charge characteristics in yeast ribosomal protein L28 interacts with secondary-structure-rich ribosomal RNA. The charge states and roles of histidine sidechains are unknown in the L28-RNA interactions, although in principle electrostatic interactions could lead to protonation.

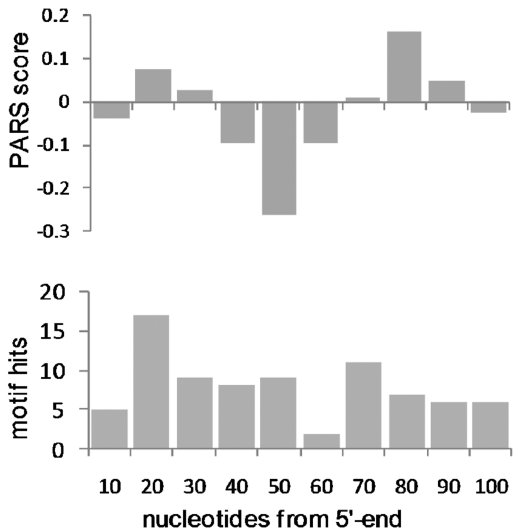

Figure 6. PARS scores and locations of a subset of RBP motifs in the 5′-UTRs of genes translationally up-regulated in the *caf20Δ* mutant. The lower panel shows the number of hits for the subset of RBPs with motifs derived in earlier work (27,28). The x-axis gives the upper nucleotide for each bin (e.g. 10 is 1–10). The upper panel gives the PARS scores, averaged over nucleotides within each motif hit and over all motif hits in a bin, also matching the nucleotide bin coordinate shown in the lower panel.

Positive charge runs and mRNA secondary structure in the context of translational regulation

Positively charged, predicted disordered regions, contribute to mRNA binding for eIF4G (19). GO analysis shows that such runs are enriched in nucleic acid-binding proteins. They are also present in the yeast 4E-BPs, where we suggest they could contribute to mRNA selectivity when coupled with mRNA secondary structure, since Caf20p has the largest positively charged region and average secondary structure is higher in the 5′-UTRs of genes up-regulated in the *caf20Δ* mutant. Studies of protein–mRNA interactions in translational control generally centre on sequence specific binding proteins, often with additional protein–protein interactions (27). The 4E-BP deletion mutant sets of up- and down-regulated genes have been examined previously in the context of enrichment for RBPs (24). In the current analysis, we were interested in the locations and PARS scores for predicted RBP sites in the 5′-UTRs. RNA motifs associated with known RBPs (27,28) were identified within the 5′-UTRs of mRNAs up-regulated in the *caf20Δ* mutant (Figure 6). RBP motifs are distributed throughout the

100 nts, shown from the 5′ end. The lower number of 5′-UTRs available in the data set as length increases will introduce a bias towards locations closer to the 5′ end. Overall, it appears that there is no clear tendency for the binding sites of this set of RBPs to lie towards the 5′ end. By contrast, the increase in PARS profile for this set of 5′-UTRs, relative to the transcriptome average (Figure 1a), is particularly apparent towards the 5′ end. In addition, the profile of PARS scores for motif hits is generally negative. Whilst it partly follows the PARS profile for *caf20Δ* up-regulated genes (e.g. with a dip around nucleotides 40–50, Figure 1a), this similarity is superposed on a different absolute scale (positive PARS scores in Figure 1a, largely negative in Figure 6a).

SUMMARY AND PERSPECTIVE

High throughput approaches are illuminating the field of post-transcriptional control, with a rich structure of RBP protein sites in mRNAs becoming apparent (28), analogous to the complexity of transcription factor binding sites in the genome. We find that regions of predicted disorder, relatively high in positive charge, are enriched in nucleic acid-binding proteins. Such positive charge runs are evident, but with variations, in the two yeast 4E-BPs that have been studied in respect of their mRNA selectivity (24). Subsets of mRNAs associated with these 4E-BPs have different 5′-UTR secondary structure properties. It is suggested that the degree of mRNA secondary structure could modulate interaction with positive charge runs in regions of protein disorder, and that the consequent variation could complement other mechanisms, in establishing mRNA selectivity of the 4E-BPs. The suggestion that positively charged runs in the 4E-BPs contribute to mRNA binding is consistent with earlier reports for positively charged, predicted disordered regions in eIF4G (19), and with the role of the RNA1 region in stabilising eIF4G-mRNA association (20).

To investigate whether the high average 5′-UTR secondary structures for *caf20Δ* up-regulated genes are associated with RBPs, we scanned the 5′-UTRs for known RBP-binding motifs. No clear association between the higher 5′-UTR secondary structure and RBPs was found. However, many RBPs remain uncharacterized in detail. Therefore, this analysis does not exclude a link between motif-specific RBPs, the mRNA secondary structure difference noted here (Figure 1a), and translational regulation. Indeed, translational control uncovered by the 4E-BP deletion mutants has been linked with the 3′-UTR binding PUF family of RBPs (24), and on a wider scale RBP target motifs are associated with translational regulation (28). Nevertheless, this work leads to the hypothesis that interactions beyond those encoded in specific RNA motif—RBP pairings could play a role in translational control.

As with many areas involving intrinsically unstructured proteins, lack of structural visualisation makes it difficult to characterize interactions in detail. Modelling of polyelectrostatic phenomena in biology (40,41), in combination with physico-chemical models for charged polymer interactions, will be required for a theoretical understanding of the binding phenomena proposed in this work. In simulation studies of entangled, linear, polyelectrolytes, it appears that oppositely charged molecules associate through wrapped intermediates (42). If this were the case for natively unstructured protein regions and mRNA, then the degree of mRNA secondary structure would influence binding. Wrapped chain intermediates are also evident in Brownian dynamics simulations of complexation between oppositely charged linear polymers with bond length asymmetry (i.e. differing charge densities) (43).

It is proposed that a relatively high positive charge run in Caf20p could be suited to interacting with mRNA that is relatively rich in secondary structure. The structure of a similar positively charged region in ribosomal protein L28 shows interactions with ribosomal RNA. Interestingly, longer 5′-UTRs (as well as higher PARS scores) associate with genes up-regulated in the *caf20Δ* mutant. Longer 5′-UTRs could lead to greater separation between paired bases, more constrained mRNA structure in 3D, and possibly higher mRNA negative charge density.

It has been suggested that auxiliary domains in eukaryotic RBPs provide relatively non-specific, preliminary protein–RNA interactions, followed by migration to form a high affinity, RNA sequence-specific interaction with a separate RNA-binding domain (44). This is similar to the process by which transcription factors are thought to use non-specific binding to facilitate intramolecular sliding and location of specific binding sites (45). Typical of proteins containing RNA-binding auxiliary domains is the SR protein family, with high affinity RNA recognition motifs and lower affinity arginine, serine-rich (RS) domains (21). The 4E-BPs and eIF4G also link to mRNA via high affinity (eIF4E-mediated) interactions. Weaker RNA-binding sites are present in eIF4G (19) and are proposed here for the 4E-BPs. Analysis of positive charge runs seems to be a more general method for identifying possible RNA binding regions than assignment of a particular (e.g. RS) domain and may therefore be useful as a tool to aid the study of auxiliary nucleic acid-binding domains.

ACKNOWLEDGEMENTS

The authors thank Graham Pavitt and Simon Hubbard for discussions and the anonymous referees for their comments.

FUNDING

Funding for open access charge: UK Biotechnology and Biological Sciences Research Council (PhD studentship award to A.C.).

Conflict of interest statement. None declared.

REFERENCES

1. Sonenberg,N. and Hinnebusch,A.G. (2009) Regulation of translation initiation in eukaryotes: mechanisms and biological targets. *Cell*, **136**, 731–745.

2. Jackson,R.J., Hellen,C.U. and Pestova,T.V. (2010) The mechanism of eukaryotic translation initiation and principles of its regulation. *Nat. Rev. Mol. Cell Biol.*, **11**, 113–127.
3. Pestova,T.V. and Kolupaeva,V.G. (2002) The roles of individual eukaryotic translation initiation factors in ribosomal scanning and initiation codon selection. *Genes Dev.*, **16**, 2906–2922.
4. Raught,B. and Gingras,A.C. (1999) eIF4E activity is regulated at multiple levels. *Int. J. Biochem. Cell Biol.*, **31**, 43–57.
5. Park,E.H., Zhang,F., Warringer,J., Sunnerhagen,P. and Hinnebusch,A.G. (2011) Depletion of eIF4G from yeast cells narrows the range of translational efficiencies genome-wide. *BMC Genomics*, **12**, 68.
6. Clarkson,B.K., Gilbert,W.V. and Doudna,J.A. (2010) Functional overlap between eIF4G isoforms in Saccharomyces cerevisiae. *PLoS One*, **5**, e9114.
7. Wilkie,G.S., Dickson,K.S. and Gray,N.K. (2003) Regulation of mRNA translation by 5′- and 3′-UTR-binding factors. *Trends Biochem. Sci.*, **28**, 182–188.
8. Mignone,F., Gissi,C., Liuni,S. and Pesole,G. (2002) Untranslated regions of mRNAs. *Genome Biol.*, **3**, REVIEWS0004.
9. Lawless,C., Pearson,R.D., Selley,J.N., Smirnova,J.B., Grant,C.M., Ashe,M.P., Pavitt,G.D. and Hubbard,S.J. (2009) Upstream sequence elements direct post-transcriptional regulation of gene expression under stress conditions in yeast. *BMC Genomics*, **10**, 7.
10. Ganoza,M.C. and Louis,B.G. (1994) Potential secondary structure at the translational start domain of eukaryotic and prokaryotic mRNAs. *Biochimie*, **76**, 428–439.
11. Robbins-Pianka,A., Rice,M.D. and Weir,M.P. (2010) The mRNA landscape at yeast translation initiation sites. *Bioinformatics*, **26**, 2651–2655.
12. Gu,W., Zhou,T. and Wilke,C.O. (2010) A universal trend of reduced mRNA stability near the translation-initiation site in prokaryotes and eukaryotes. *PLoS Comput. Biol.*, **6**, e1000664.
13. van der Velden,A.W. and Thomas,A.A. (1999) The role of the 5' untranslated region of an mRNA in translation regulation during development. *Int. J. Biochem. Cell Biol.*, **31**, 87–106.
14. Ringner,M. and Krogh,M. (2005) Folding free energies of 5'-UTRs impact post-transcriptional regulation on a genomic scale in yeast. *PLoS Comput. Biol.*, **1**, e72.
15. Sonenberg,N. and Dever,T.E. (2003) Eukaryotic translation initiation factors and regulators. *Curr. Opin. Struct. Biol.*, **13**, 56–63.
16. von der Haar,T., Gross,J.D., Wagner,G. and McCarthy,J.E. (2004) The mRNA cap-binding protein eIF4E in post-transcriptional gene expression. *Nat. Struct. Mol. Biol.*, **11**, 503–511.
17. Lorsch,J.R. and Dever,T.E. (2010) Molecular view of 43 S complex formation and start site selection in eukaryotic translation initiation. *J. Biol. Chem.*, **285**, 21203–21207.
18. von der Haar,T., Oku,Y., Ptushkina,M., Moerke,N., Wagner,G., Gross,J.D. and McCarthy,J.E. (2006) Folding transitions during assembly of the eukaryotic mRNA cap-binding complex. *J. Mol. Biol.*, **356**, 982–992.
19. Berset,C., Zurbriggen,A., Djafarzadeh,S., Altmann,M. and Trachsel,H. (2003) RNA-binding activity of translation initiation factor eIF4G1 from Saccharomyces cerevisiae. *RNA*, **9**, 871–880.
20. Park,E.H., Walker,S.E., Lee,J.M., Rothenburg,S., Lorsch,J.R. and Hinnebusch,A.G. (2011) Multiple elements in the eIF4G1 N-terminus promote assembly of eIF4G1*PABP mRNPs in vivo. *EMBO J.*, **30**, 302–316.
21. Shepard,P.J. and Hertel,K.J. (2009) The SR protein family. *Genome Biol.*, **10**, 242.
22. Nikolakaki,E., Drosou,V., Sanidas,I., Peidis,P., Papamarcaki,T., Iakoucheva,L.M. and Giannakouros,T. (2008) RNA association or phosphorylation of the RS domain prevents aggregation of RS domain-containing proteins. *Biochim. Biophys. Acta*, **1780**, 214–225.
23. Magee,J. and Warwicker,J. (2005) Simulation of non-specific protein-mRNA interactions. *Nucleic Acids Res.*, **33**, 6694–6699.
24. Cridge,A.G., Castelli,L.M., Smirnova,J.B., Selley,J.N., Rowe,W., Hubbard,S.J., McCarthy,J.E., Ashe,M.P., Grant,C.M. and Pavitt,G.D. (2010) Identifying eIF4E-binding protein translationally-controlled transcripts reveals links to mRNAs bound by specific PUF proteins. *Nucleic Acids Res.*, **38**, 8039–8050.
25. Kertesz,M., Wan,Y., Mazor,E., Rinn,J.L., Nutter,R.C., Chang,H.Y. and Segal,E. (2010) Genome-wide measurement of RNA secondary structure in yeast. *Nature*, **467**, 103–107.
26. Nagalakshmi,U., Wang,Z., Waern,K., Shou,C., Raha,D., Gerstein,M. and Snyder,M. (2008) The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science*, **320**, 1344–1349.
27. Hogan,D.J., Riordan,D.P., Gerber,A.P., Herschlag,D. and Brown,P.O. (2008) Diverse RNA-binding proteins interact with functionally related sets of RNAs, suggesting an extensive regulatory system. *PLoS Biol.*, **6**, e255.
28. Riordan,D.P., Herschlag,D. and Brown,P.O. (2011) Identification of RNA recognition elements in the Saccharomyces cerevisiae transcriptome. *Nucleic Acids Res.*, **39**, 1501–1509.
29. Linding,R., Russell,R.B., Neduva,V. and Gibson,T.J. (2003) GlobPlot: exploring protein sequences for globularity and disorder. *Nucleic Acids Res.*, **31**, 3701–3708.
30. Prilusky,J., Felder,C.E., Zeev-Ben-Mordehai,T., Rydberg,E.H., Man,O., Beckmann,J.S., Silman,I. and Sussman,J.L. (2005) FoldIndex: a simple tool to predict whether a given protein sequence is intrinsically unfolded. *Bioinformatics*, **21**, 3435–3438.
31. Warwicker,J. (1999) Simplified methods for pKa and acid pH-dependent stability estimation in proteins: removing dielectric and counterion boundaries. *Protein Sci.*, **8**, 418–425.
32. Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P.E. (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242.
33. Altschul,S.F. and Koonin,E.V. (1998) Iterated profile searches with PSI-BLAST: a tool for discovery in protein databases. *Trends Biochem. Sci.*, **23**, 444–447.
34. Ashkenazy,H., Erez,E., Martz,E., Pupko,T. and Ben-Tal,N. (2010) ConSurf 2010: calculating evolutionary conservation in sequence and structure of proteins and nucleic acids. *Nucleic Acids Res.*, **38**, W529–W533.
35. Boyle,E.I., Weng,S., Gollub,J., Jin,H., Botstein,D., Cherry,J.M. and Sherlock,G. (2004) GO::TermFinder: open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes. *Bioinformatics*, **20**, 3710–3715.
36. Shabalina,S.A., Ogurtsov,A.Y., Rogozin,I.B., Koonin,E.V. and Lipman,D.J. (2004) Comparative analysis of orthologous eukaryotic mRNAs: potential hidden functional signals. *Nucleic Acids Res.*, **32**, 1774–1782.
37. Garfinkel,D.J., Nyswaner,K.M., Stefanisko,K.M., Chang,C. and Moore,S.P. (2005) Ty1 copy number dynamics in Saccharomyces. *Genetics*, **169**, 1845–1857.
38. Sandmeyer,S.B. and Clemens,K.A. (2010) Function of a retrotransposon nucleocapsid protein. *RNA Biol.*, **7**, 642–654.
39. Ben-Shem,A., Garreau de Loubresse,N., Melnikov,S., Jenner,L., Yusupova,G. and Yusupov,M. (2011) The structure of the eukaryotic ribosome at 3.0 A resolution. *Science*, **334**, 1524–1529.
40. Borg,M., Mittag,T., Pawson,T., Tyers,M., Forman-Kay,J.D. and Chan,H.S. (2007) Polyelectrostatic interactions of disordered ligands suggest a physical basis for ultrasensitivity. *Proc. Natl Acad. Sci. USA*, **104**, 9650–9655.
41. Mittag,T., Kay,L.E. and Forman-Kay,J.D. (2010) Protein dynamics and conformational disorder in molecular recognition. *J. Mol. Recognit.*, **23**, 105–116.
42. Winkler,R.G., Steinhauser,M.O. and Reineker,P. (2002) Complex formation in systems of oppositely charged polyelectrolytes: a molecular dynamics simulation study. *Phys. Rev.*, **66**, 021802.
43. Trejo-Ramos,M.A., Tristan,F., Menchaca,J.L., Perez,E. and Chavez-Paez,M. (2007) Structure of polyelectrolyte complexes by Brownian dynamics simulation: effects of the bond length asymmetry of the polyelectrolytes. *J. Chem. Phys.*, **126**, 014901.
44. Biamonti,G. and Riva,S. (1994) New insights into the auxiliary domains of eukaryotic RNA binding proteins. *FEBS Lett.*, **340**, 1–8.
45. Clore,G.M., Tang,C. and Iwahara,J. (2007) Elucidating transient macromolecular interactions using paramagnetic relaxation enhancement. *Curr. Opin. Struct. Biol.*, **17**, 603–616.
46. Gross,J.D., Moerke,N.J., von der Haar,T., Lugovskoy,A.A., Sachs,A.B., McCarthy,J.E. and Wagner,G. (2003) Ribosome loading onto the mRNA cap is driven by conformational coupling between eIF4G and eIF4E. *Cell*, **115**, 739–750.