Acta Biochimica Polonica

Vol. 54 No. 1/2007, 55–61

on-line at: www.actabp.pl

Regular paper

Prediction of the structure of the common perimitochondrial localization signal of nuclear transcripts in yeast

Radoslaw K. Ejsmont[1], Pawel Golik[1,2] and Piotr P. Stepien[1,2]✉

[1]*Institute of Genetics and Biotechnology, University of Warsaw, Warszawa, Poland;* [2]*Institute of Biochemistry and Biophysics PAS, Warszawa, Poland*

Received: 11 January, 2007; revised: 06 March, 2007; accepted: 16 March, 2007
available on-line: 20 March, 2007

Many nuclear genes encoding mitochondrial proteins require specific localization of their mRNAs to the vicinity of mitochondria for proper expression. Studies in *Saccharomyces cerevisiae* have shown that the *cis*-acting signal responsible for subcellular localization of mRNAs is localized in the 3′ UTR of the transcript. In this paper we present an *in silico* approach for prediction of a common perimitochondrial localization signal of nuclear transcripts encoding mitochondrial proteins. We computed a consensus structure for this signal by comparison of 3′ UTR models for about 3000 yeast transcripts with known localization. Our studies show a short stem-loop structure which appears in most mRNAs localized to the vicinity of mitochondria. The degree of similarity of a given 3′ UTR to our consensus structure strongly correlates with experimentally determined perimitochondrial localization of the mRNA, therefore we believe that the structure we predicted acts as a subcellular localization signal. Since our algorithm operates on structures, it seems to be more reliable than sequence-based algorithms. The good predictive value of our model is supported by statistical analysis.

Keywords: mitochondria, RNA structure, RNA trafficking, bioinformatics

INTRODUCTION

Subcellular localization of a given mRNA may play a crucial role in correct functioning of the respective protein in the cell. Some proteins require mRNA localization for their expression in specific subcellular compartments, for example those involved in embryo development, neural activity and in mitochondrial biogenesis. Many of the about 1000 mitochondrial proteins encoded by the nuclear genome are synthesized in a process that strictly requires localization of their transcripts in the vicinity of subcellular structures, such as mitochondria (Jansen *et al*., 2001).

Studies in *Saccharomyces cerevisiae* have proved that the *cis*-acting signal responsible for mRNA localization to the vicinity of mitochondria is localized in the 5′ UTR and/or 3′ UTR of the *ATM1* gene. Perturbation of this signal can lead to incorrect localization, which results in respiratory dysfunction, despite the fact that the ORF has not been changed (Corral-Debrinski *et al*., 2000). In addition, results of a genome-wide microarray assay experiment have shown that a signal in the 3′ UTR is responsible for mitochondrial localization of dozens of nuclear mRNAs encoding mitochondrial proteins. It has also been reported that most proteins translated from mRNAs with perimitochondrial localization are of a prokaryotic origin (Marc *et al*., 2002). Following subsequent studies, a 3′ UTR stem-loop structure responsible for sorting the *ATP2* mRNA to the vicinity of mitochondria has been predicted (Margeot *et al*., 2002). The role of the 3′ UTR signal responsible for mRNA distribution between free and mitochondria-bound polysomes has also been shown to be conserved in the eukaryotic world (Sylvestre *et al*., 2003).

Those experiments suggest that a 3′ UTR signal, probably a stem-loop structure, could be a universal tag marking mRNAs for transport to the

✉Correspondence author: Piotr P. Stepien, Institute of Genetics and Biotechnology, University of Warsaw, Pawińskiego 5a, 02-106 Warszawa, Poland; tel.: (48 22) 592 2240; fax: (48 22) 658 4176; e-mail: stepien@ibb.waw.pl

vicinity of mitochondria. In this paper we present an approach to predict *in silico* a structure that is common to most transcripts that represent a high Mitochondrial localization ratio (MLR), by analyzing data from the experiment performed by Marc *et al.* (2002).

METHODS

Sequences and databases. All yeast genome sequences were downloaded from the *Saccharomyces cerevisiae* Genome Database (SGD) (Cherry *et al.*, 1997). The database containing fungal 3′ UTR sequences (UTRdb) was downloaded from Internet Resources for UTR analysis (Pesole *et al.*, 2002). Localization data for nuclear transcripts encoding mitochondrial proteins were downloaded from the LGM Mitochondria Microarray Project Web site (Marc *et al.*, 2002).

We have downloaded sequences of 3050 genes out of 3106 analyzed in (Marc *et al.*, 2002). Sequences for 56 genes were unavailable, thus not analyzed.

RNA sequence and structure prediction. The 3′ UTR sequences were predicted as described in the Results section and in Fig. 1. We used ClustalW (Higgins *et al.*, 1994) to make a global alignment and WUstl-BLAST (Gish *et al.*, 1996) for local alignment and sequence screening. Structure prediction was performed with mFold (Zuker *et al.*, 2003). Structural alignments were done using RNAdistance from the ViennaRNA package (Hofacker, 2003).

Data analysis. We tested the quality of both sequences and structures we have predicted. Since the –log(E-value) indicates the likelihood that the predicted sequence is aligned correctly and Gibbs' free energy (ΔG) presents the stability of the predicted structures, these values were used to measure the quality. We also compared the length and the GC-content of the 3′ UTR sequences in our database.

All of the predicted structures had some common elements. We applied statistical methods to decide if differences between groups of structures with different MLR present enough diversity to consider our template as a common element of mRNAs that are transported to the vicinity of mitochondria. Thus we computed average ScoreD for groups of structures with similar MLR values and evaluated the correlation between ScoreD and MLR by linear regression. Moreover, to make sure that a statistically significant difference exists between groups with different MLRs, we applied a test comparing ScoreD for subsets with a high (>90) and low (<10) MLR value.

RESULTS

Identification of the 3′ UTR sequences

Since yeast EST (see Sequences and databases) databases do not contain the majority of transcripts, we had to predict most of the 3′ UTR sequences from genomic sequences. For each analyzed gene a sequence of 2000 nucleotides downstream from the STOP codon of the respective ORF was prepared. We called it the 2 kb tail. It was used as a BLAST query against the UTRdb (Pesole *et al.*, 2002). We extracted from the UTRdb sequences with the highest homology to the query and globally realigned them to the query sequence using ClustalW. The existence of the poly-A tails in the query sequences could disturb the alignment, therefore they were truncated. Alignment files were parsed by our software and 3′ UTR sequences, beginning at the first nucleotide of the 2 kb tail sequence and ending at the last matched nucleotide of the alignment were extracted. Both 2 kb tail and 3′ UTR sequences for each analyzed gene were placed together with the genomic and coding sequences in an SQL database. The whole sequence extraction procedure is illustrated in Fig. 1.

From the 3050 sequences downloaded from SGD, we were able to predict 3′ UTR sequences for 2953 genes. The shortest predicted sequence had the length of 45 bp, whereas the longest 2 kb. The average sequence length was 1103 bp. The distribution of sequence lengths across the database was similar for subsets with different MLRs and very close to the average for the whole database. We observed a similar pattern in respect to the GC-content in the predicted sequences, which was approx. 34.25%. The average –log(E-value), representing the probabil-

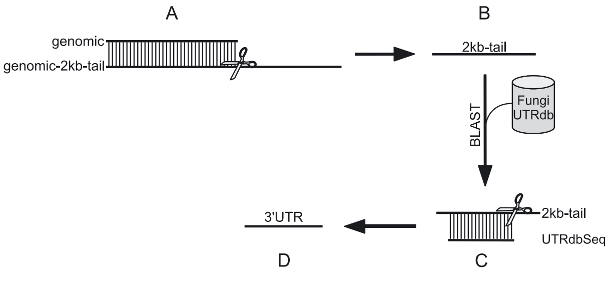

Figure 1. 3′ UTR sequence extraction procedure.
Genomic sequence (genomic) is aligned using ClustalW with the genomic sequence containing 2 kb downstream region (genomic-2kb-tail) and the downstream region is extracted (**A**). Extracted sequence (2kb-tail) is used as a BLAST query against Fungi UTRdb and the highest scoring sequence is extracted (**B**). The sequence from Fungi UTRdb is aligned using ClustalW with the 2 kb-tail sequence (**C**). 3′ UTR sequence is extracted from alignment. It begins with first nucleotide of 2 kb-tail sequence and ends at the last matched nucleotide in alignment (**D**).

ity of a successful alignment, was for our database 33.54, indicating that it was of a high quality.

Identification of the template for transport signal prediction

In order to predict a common signal for perimitochondrial localization of mRNA we decided to find one model 3′ UTR structure (a template) based on the criteria listed below, and to look for its occurrence in other 3′ UTR structures. If the structure we have chosen was a real mitochondrial localization signal, there should be a correlation between the experimental MLR values from Marc *et al*. (2002) and the similarity of a given mRNA structure to our template.

We used the following criteria in our search for the template:

— The structure should be small, less than 200 nucleotides long, since such small structures were reported previously as controlling mRNA localization (Margeot *et al*., 2002);

— The MLR value for mRNA containing this structure should be high, not less than 95;

— The protein encoded by the mRNA containing the mitochondrial localization signal should have prokaryotic homologs (Marc *et al*., 2002).

— It would be of advantage if a gene containing our template 3′ UTR had a well-proven mitochondrial function or would be phylogenetically connected with the mitochondrial genome.

The sequence that satisfied the majority of these conditions was the 3′ UTR of the YJL225C gene, with the MLR = 99. This gene is localized on yeast chromosome X and encodes a protein with a helicase activity (Yamada *et al*., 1998). It has a CDS of 5277 bp and a short intron of 388 bp. The 3′ UTR of YJL225C is very short and consists of only 72 bp.

The sequence of YJL225C is almost identical to the 3SCE000226 sequence from UTRdb and presents partial homology with RecG helicase from *Chlorobium tepidum* and BH1607 helicase from *Bacillus halodurans*. It is very likely that YJL225C is of mitochondrial origin, since the sequence downstream from the CDS contains a fragment identical to a part of the bI4-intron of the cytochrome *b* gene. No experimental data on the localization of the gene product have been reported in the literature. On the other hand, the MitoProtII prediction showed a rather low probability (13.70%) that this protein is localized in mitochondria.

We performed structure prediction for the 3′ UTR sequence of our template using mFold and got a 72 bp stem-loop structure with five stems and five loops (Fig. 2). Δ*G* of the predicted structure was –8.98 kcal/mol and the average pairing energy was 125 cal/mol.

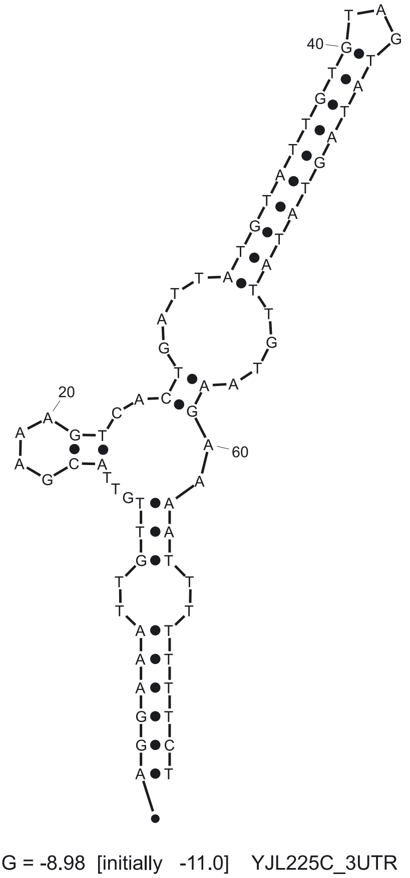

Figure 2. Putative structure of the YJL225C 3′ UTR — the predicted perimitochondrial localization signal.

RNA structure prediction

The next step was to predict the structure of the template, as well as structures of other 3′ UTR sequences in our database using mFold (Zuker, 2003). The predicted 2D structures were then converted to bracket notation and placed in our database. All of the predicted structures were aligned with the template using RNA distance from the ViennaRNA software package (Hofacker *et al*., 2003) and assigned scores describing the distance between the analyzed structure and the template. The following scoring was used:

— ScoreA — number of positions in the analyzed structure **differing** from those in the template divided by the analyzed structure's length;

— ScoreB — number of positions in the template **absent** in the analyzed structure, divided by 0.05% of the template's length (factor based on normalization to the average value for the whole dataset);

— ScoreC — number of inner (i.e., those flanked by already aligned fragments) positions in the analyzed structure **absent** in the template, divided by the analyzed structure's length;

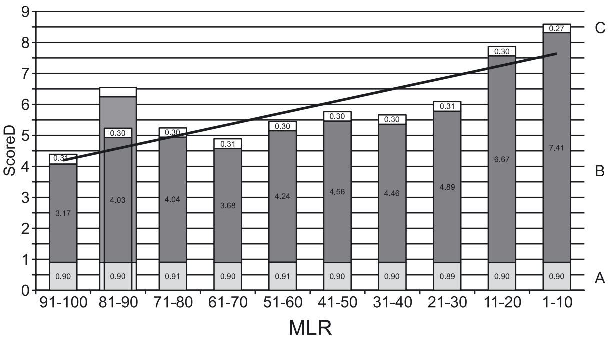

Figure 3. Correlation between MLR value and ScoreD. The light grey parts of bars represent ScoreA, dark grey — ScoreB and the white ones — ScoreC. The total height of bar represents ScoreD. The number of transcripts in each group is shown in Fig. 4. There are two bars shown representing ScoreD for transcripts with MLR from 81 to 90. This was done due to a few suprisingly high ScoreD values for some transcripts. The list of these transcripts is presented in Table 2. The narrower bar shows ScoreD calculated excluding these values. The line represents growing trend calculated by linear regression (excluding the suprisingly high values in MLR 81–90 group) with regression error of 0.77 and r^2 of 0.79.

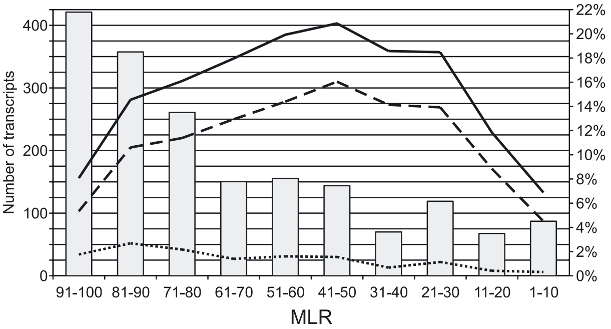

Figure 4. Number of genes in each analyzed MLR group. Solid line represents total number of genes, dashed line represents number of genes encoding proteins with known localization, dotted line applies to genes encoding proteins with known mitochondrial localization (data based on GFP assays (Huh *et al.*, 2003)). Bars show percentage of genes encoding proteins with known mitochondrial localization in each group.

— ScoreD — the sum of ScoreA, ScoreB and Score C. The **negative** (ScoreD) of this value represents the similarity of the analyzed structure to that of the template.

The computed data were put into the database for further analysis.

Structure prediction for 2953 tested yeast 3′ UTR sequences in our database produced 78705 structures, approx. 27 structures per each 3′ UTR sequence. Each structure had ScoreA, ScoreB, ScoreC, ScoreD and ΔG values assigned, based on the results of our computations.

We chose to work on structures with the lowest ΔG, one ΔG value per each gene (5961 structures, about two structures per sequence), since the others presented random ScoreD distribution with respect to MLR (results not shown). The average energy of the selected subset was –228 kcal/mol, average pairing energy was 192 cal/mol.

We computed average ScoreA, ScoreB, ScoreC and ScoreD for groups with different MLR values and found a linear correlation between ScoreD and MLR values. High ScoreD value corresponds to a

Table 1. Genes with low (about 1) ScoreD for which products do localize in mitochondria (based on GFP assays, Huh *et al.*, 2003).

ATP2 has also been presented for reference, although it has a higher ScoreD.

Locus	Gene	ScoreD	MLR	Protein localization (GFP)
YBR026C	*MRF1*	0.990996	97	mitochondrion
YOR108W		1.00549	97	mitochondrion
YGR183C	*QCR9*	1.00567	49	mitochondrion
YHL018W		1.00881	54	mitochondrion
YCL064C	*CHA1*	1.01279	90	mitochondrion
YBR263W	*SHM1*	1.01824	92	mitochondrion
YLL027W	*ISA1*	1.01828	43	mitochondrion
YHR067W		1.02104	60	mitochondrion
YFL027C		1.02274	47	mitochondrion
YDR234W	*LYS4*	1.02293	56	mitochondrion
YGR012W		1.02564	50	mitochondrion
YBL090W	*MRP21*	1.02641	43	mitochondrion
YCR083W	*TRX3*	1.02714	27	mitochondrion
YKR049C		1.03002	16	mitochondrion
YDL027C		1.03222	85	mitochondrion
YJR121W	*ATP2*	1.45952	98	mitochondrion

Table 2. Genes with extremely high (>50) ScoreD and MLR from 81 to 90.

Localization data based on GFP assays (Huh ***et al.***, 2003).

Locus	Gene	ScoreD	MLR	Protein localization (GFP)
YBL022C	*PIM1*	81.9598	89	mitochondrion
YBR296C	*PHO89*	59.4176	85	endoplasmic reticulum (ER)
YCL038C	*AUT4*	85.2561	88	ER
YER077C		65.8106	81	ER, mitochondrion

low similarity to the template, and correlates with low MLR values, low ScoreD represents structures with high MLR. Correlation was determined using linear regression with $r^2=0.77$ and a regression error of 0.66. See Fig. 3 for details.

In total, 2953 yeast 3′ UTRs were analyzed. Structures for 2302 contained some fragments of the YJL225C structure, we assumed therefore that this structure is as close to the consensus as possible.

Template verification

Since the biological criteria used for the best template selection were arbitrary, we had to check if the template selected on the basis of those criteria was indeed the best template possible. Therefore we selected 192 possible templates with MLR values greater than 88 and calculated correlation coefficients between ScoreD calculated basing on those templates and MLRs for the tested dataset of 2953 yeast 3′ UTRs. The results of these computations are plotted on Fig. 5. The strongest negative correlation between ScoreD and MLR exists for YJL225C, the template which we preselected.

Statistical analysis of computed data

In order to verify if the structure found was good enough to be considered as a consensus for all transcripts that follow the perimitochondrial localization pathway, we had to apply statistical methods that would show whether the determined correlation was statistically significant. The statistical test showed that the suggested dependency was statistically significant.

DISCUSSION

The obtained data suggest that the YJL225C 3′ UTR is a good consensus structure for the perimitochondrial localization signal. It is short (probably due to subtelomeric localization of the gene), has a high MLR value (Marc *et al.*, 2002), high AU content, has bacterial homologs and due to homology with mitochondrial DNA (see "Identification of the template for transport signal prediction"), is related to the mitochondrial genome. The protein encoded by YJL225C does not contain a mitochondrial import signal, however, there are other examples of

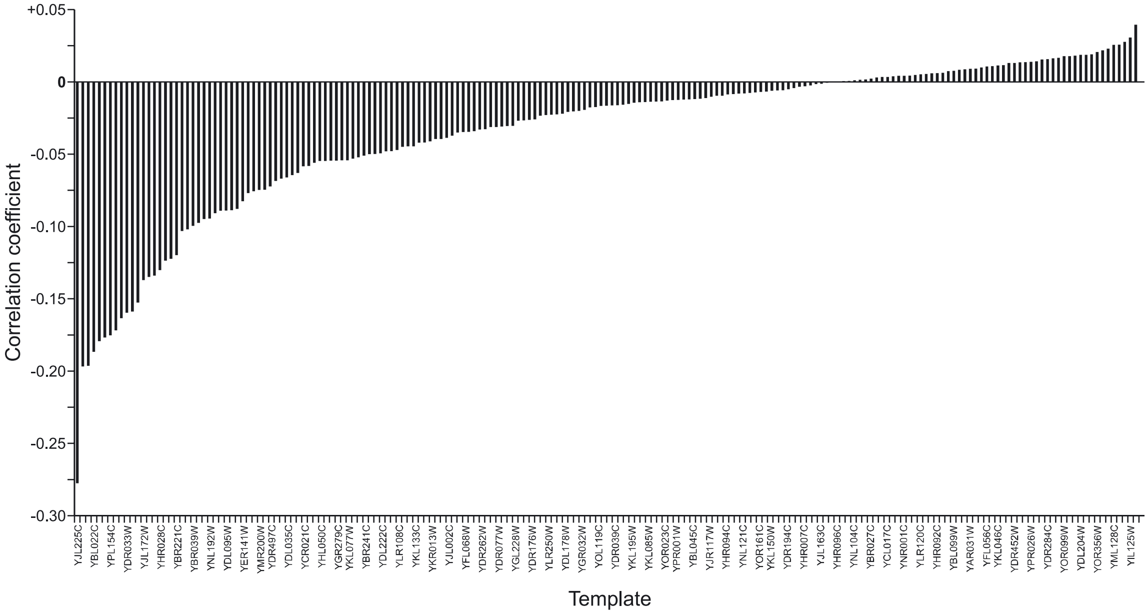

Figure 5. A comparison of correlation coefficients for different 3′ UTRs used as templates for the common perimitochondrial localization signal.

YJL225C presents the strongest negative correlation between ScoreD and MLR.

proteins that do localize in mitochondria but do not contain an import sequence. However, a mitochondrial function of YJL225C is not strictly required for this protein. It could have been lost during evolution. The strong perimitochondrial localization signal in YJL225C 3′ UTR could also have been acquired by recombination.

The modeled structure for the YJL225C 3′ UTR sequence is similar to the previously predicted structure of *ATP2* 3′ UTR, a transcript proved to localize to the vicinity of mitochondria (Margeot *et al*., 2002). The low energy of the modeled structure suggests that it is stable in the cellular environment. Fragments of the YJL225C 3′ UTR appear in most of the analyzed structures, furthermore, the structures with high (>90) MLR values contain nearly the entire YJL225C structure, whereas in those with low (<10) MLR large fragments of this structure are missing.

Our analysis shows that the 3′ UTR length has no influence on perimitochondrial localization. We observed that the 3′ UTR regions of many genes overlap with 5′ UTR regions or even CDS of downstream genes which is normal in the highly compact yeast genome as well as in genomes of other primitive Eucaryota. The great length of many sequences suggested that the assumed threshold of 2 kb could miss some sequences. In fact, we found 115 sequences with a length of 2 kb, but since they were a minor (less than 4%) part of the database, we ignored them.

We observed a high AU content in the predicted 3′ UTR sequences, which is typical for untranslated regions of mRNA. Most regulatory regions are described as AU-rich sequences, especially those connected with transcription termination and polyadenylation (Grafi *et al*., 1993; Graber *et al*., 1999; Legendre *et al*., 2003; Caballero *et al*., 2004). Our analysis has shown that the GC content does not influence the perimitochondrial localization pathway.

Our analysis revealed that neither the average pairing energy nor total energy affects perimitochondrial localization. This suggests that in all cases only a small fragment of the 3′ UTR is responsible for targeting the mRNAs to the vicinity of mitochondria. In addition, it appears that other signals in the long 3′ UTR do not interfere with the perimitochondrial localization signal.

CONCLUSIONS

In this paper we predicted *in silico* a structure of 3′ UTR responsible for perimitochondrial localization of cytoplasmic yeast mRNAs. We have analyzed almost half of the yeast transcriptome for which the MLR values were determined.

Our method is based on structure analysis that, compared with sequence based algorithms (as the one used for example by Jacobs Anderson and Parker (Jacobs *et al*., 2000)), should give more reliable results. To give an example, *ATP2* mRNA known to localize in the vicinity of mitochondria does not contain the CYTGTAAATA element described in (Jacobs *et al*., 2000), but does contain a structure similar to that of the YJL225C 3′ UTR. Of course there is a group of 3′ UTRs that do contain the CYTGTAAATA element and a structure similar to that of the YJL225C 3′ UTR (four genes with ScoreD <1, 34 genes with ScoreD <1.5).

The model we developed has a potentially high predictive value for perimitochondrial localization of transcripts with unknown MLR due to the strong correlation between ScoreD and MLR that has been estimated by $r^2=0.77$ of linear regression (Fig. 3).

To summarize, we have shown that mRNAs following the perimitochondrial localization pathway in yeast contain a common structural signal, similar to the one found in the YJL225C 3′ UTR. The data acquired from our computation strongly correlate with empirical results from (Marc *et al*., 2002). When we confirm our results *in vivo* we will be able to create an algorithm for fast *in silico* prediction of perimitochondrial localization of any yeast mRNA sequence.

Acknowledgements

This work was supported by the Committee for Scientific Research (KBN) grant PBZ-KBN 091/P05/2003. R.E. was supported by a Stipend of the Ministry of Education and Science.

We would like to thank Professor Ewa Bartnik for critical reading of the manuscript.

REFERENCES

Caballero JJ, Giron MD, Vargas AM, Sevillano N, Suarez MD, Salto R (2004) AU-rich elements in the mRNA 3′-untranslated region of the rat receptor for advanced glycation end products and their relevance to mRNA stability. *Biochem Biophys Res Commun* **319:** 247–255.

Cherry JM, Ball C, Weng S, Juvik G, Schmidt R, Adler C, Dunn B, Dwight S, Riles L, Mortimer RK, Botstein D (1997) Genetic and physical maps of *Saccharomyces cerevisiae. Nature* **387:** 67–73.

Corral-Debrinski M, Blugeon C, Jacq C (2000) In yeast the 3′untranslated region or the presequence of *ATM1* is required for the exclusive localization of its mRNA to the vicinity of mitochondria. *Mol Cell Biol* **20:** 7881–7892.

Gish W (1996–2003) WU-BLAST http://blastwustledu.

Graber JH, Cantor CR, Mohr SC, Smith TF (1999) Genomic detection of new yeast pre-mRNA 3′-end-processing signals. *Nucleic Acids Res* **27:** 888–894.

Grafi G, Sela I, Galili G (1993) Translational regulation of human beta interferon mRNA: association of the 3′ AU-rich sequence with the poly(A) tail reduces translation efficiency *in vitro*. *Mol Cell Biol* **13:** 3487–3493.

Higgins D, Thompson J, Gibson T, Thompson JD, Higgins DG, Gibson TJ (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting position-specific gap penalties and weight matrix choice. *Nucleic Acids Res* **22:** 4673–4680.

Hofacker IL (2003) Vienna RNA secondary structure server *Nucleic Acids Res* **31:** 3429–3431.

Huh WK, Falvo JV, Gerke LC, Carroll AS, Howson RW, Weissman JS, O'Shea EK (2003) Global analysis of protein localization in budding yeast. *Nature* **425:** 686–691.

Jacobs Anderson JS, Parker R (2000) Computational identification of *cis*-acting elements affecting post-transcriptional control of gene expression in *Saccharomyces cerevisiae. Nucleic Acids Res* **28:** 1604–1617.

Jansen RP (2001) mRNA localization: message on the move. *Nat Rev Mol Cell Biol* **2:** 247–256.

Legendre M, Gautheret D (2003) Sequence determinants in human polyadenylation site selection. *BMC Genomics* **4**.

Marc P, Margeot A, Devaux F, Blugeon C, Corral-Debrinski M, Jacq C (2002) Genome-wide analysis of mRNAs targeted to yeast mitochondria. *EMBO Rep* **3:** 159–164.

Margeot A, Blugeon C, Sylvestre J, Vialette S, Jacq C, Corral-Debrinski M (2002) In *Saccharomyces cerevisiae ATP2* mRNA sorting to the vicinity of mitochondria is essential for respiratory function. *EMBO J* **21:** 6893–6904.

Pesole G, Liuni S, Grillo G, Licciulli F, Mignone F, Gissi C, Saccone C (2002) UTRdb and UTRsite: specialized databases of sequences and functional elements of 5′ and 3′ untranslated regions of eukaryotic mRNAs; Update 2002. *Nucleic Acids Res* **30:** 335–340.

Sylvestre J, Margeot A, Jacq C, Dujardin G, Corral-Debrinski M (2003) The role of the 3′untranslated region in mrna sorting to the vicinity of mitochondria is conserved from yeast to human cells. *Mol Biol Cell* **14:** 3848–3856.

Yamada M, Hayatsu N, Matsuura A, Ishikawa F (1998) Y′-Help1 a DNA helicase encoded by the yeast subtelomeric Y′ element is induced in survivors defective for telomerase. *J Biol Chem* **273:** 33360–33366.

Zuker M (2003) Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res* **31:** 3406–3415.