Examination of the tRNA Adaptation Index as a Predictor of Protein Expression Levels

Orna Man[1,2], Joel L. Sussman[1], and Yitzhak Pilpel[2]

[1] Department of Structural Biology, The Weizmann Institute of Science, Rehovot 76100, Israel
{orna.man, joel.sussman}@weizmann.ac.il
http://www.weizmann.ac.il/~joel
[2] Department of Molecular Genetics, The Weizmann Institute of Science, Rehovot 76100, Israel
pilpel@weizmann.ac.il
http://longitude.weizmann.ac.il

Abstract. Phenotypic differences between closely-related species may arise from differential expression regimes, rather than different gene complements. Knowledge of cellular protein levels across a species sample would thus be useful for the inference of the genes underlying such phenotypic differences. dos Reis et al [1] recently proposed the tRNA Adaptation Index to score the optimality of a coding sequence with respect to a species' cellular tRNA pools. As a preliminary step towards a multi-species analysis that would utilize this index, we examine in this paper its performance in predicting protein expression levels in the yeast *S. cerevisiae* and find that it likely predicts maximal potential levels of proteins. We also show that tAI profiles of genes across species carry functional information regarding the interactions between proteins.

1 Introduction

A major challenge in evolutionary research is to understand molecular and genomic causes of phenotypic divergence of species. One obvious source of difference in phenotype and life-style among related species may be differences in their gene complements. The phylogenetic profiles method [2] utilizes this concept by clustering together genes that share the same pattern of presence/absence in the genomes of a set of species. A striking example [3] is the recent identification of an entire set of genes involved in the formation of cilia - short hair-like appendages found on the surfaces of some types of cells in some organisms. The genes were identified on the basis of their presence in all (sequenced) species known to have ciliated cells and absence in all (sequenced) species known to be devoid of this sub-cellular structure.

Such a methodology, although useful, is necessarily limited to genes that are found in only a fraction of the species analyzed, and may be problematic when one would like to make inferences for closely related species. Martin et al [4] created a matrix denoting for each *E. coli* open reading frame (ORF) its conservation, in a variety of prokaryotic species, relative to the *E. coli* sequence. They suggested that clustering this matrix by genes could lead to genotype-to-phenotype associations, and could perhaps even reveal the genes responsible for specific traits. As an example they identified functions that are over-represented in genes differentiating between Gram-positive and Gram-negative bacteria.

E. Eskin et al. (Eds.): RECOMB 2005 Ws on Regulatory Genomics, LNBI 4023, pp. 107 – 118, 2007.

Recently, it has been shown that predicted levels of expression of functionally related proteins tend to co-evolve [5, 6] allowing the study of interactions between proteins present in all analyzed species. These studies utilized the Codon Adaptation Index (CAI) [7] as a predictor of protein levels. The CAI infers the optimality weights of the various codons by examining the codon usage of the coding sequences of a group of genes that are assumed to be highly expressed in the species examined, and uses these weights to judge the optimality of any coding sequence in this species with respect to translation. The underlying assumption of this method is that the coding sequences of highly expressed genes are well-adapted to the tRNA pools of the cell, so that their codon usage reflects these pools, and therefore allows for the inference of optimality scores for codons, and consequently for coding sequences. The observation that cellular tRNA pools are highly correlated with the tRNA gene copy numbers in the genome [8], allows for the inference of codon optimality scores more directly, without the need to select a group of highly expressed genes. Indeed, a recent study suggested an alternative index of translational optimality – the tRNA Adaptation Index (tAI) [1], and demonstrated its use for the inference of genome-wide translational selection. As a preliminary step towards the analysis of phenotypic divergence using tAI, we examine in this paper the utility of this index as a predictor of protein expression levels, as well as the functional content of multi-species tAI profiles.

2 Results and Discussion

2.1 tAI as an Indicator of Protein Expression Levels

The tAI predicts the level of adaptation of a coding sequence relative to the cell's tRNA pools. As a first test of the functional power of prediction of this index we examined its correspondence with genome-wide experimentally determined protein levels in *S. cerevisiae* [9]. Using the protein levels of almost 4000 *S. cerevisiae* ORFs we obtain a significant positive correlation (R=0.63 using Pearson correlation; p<1e-363) between tAI values and the corresponding log-transformed protein levels (Fig. 1A). The same analysis, using a different data set constituting 150 proteins [10], yielded similar results. Comparable, yet lower, correlations were obtained using the related indices CAI [7] (R=0.58) and FOP' [11] (R=0.57).

Significant correlations have been previously observed between CAI and mRNA levels [12], presumably due the general association of high protein levels with high transcript levels. Indeed, the correlation between the log-transformed genome-wide mRNA [13] and protein [9] levels obtained under similar conditions is highly significant with a Pearson correlation coefficient R=0.62 (p<1e-363; Fig. 1B). However, transcript levels are generally considered poor indicators of protein levels, as similar mRNA levels may be accompanied by a wide range (up to 20-fold difference) of protein levels, and *vice versa* [14]. Some of the discrepancy between mRNA and protein levels may be perhaps explained by different levels of translational control exerted on genes with a similar mRNA level. Such translational control may be manifested in the adaptation of the coding sequence of genes to the tRNA pools of the cell. Therefore, the tAI could potentially provide complementary information to mRNA levels when predicting protein levels. To examine the contribution of the tAI in the prediction of

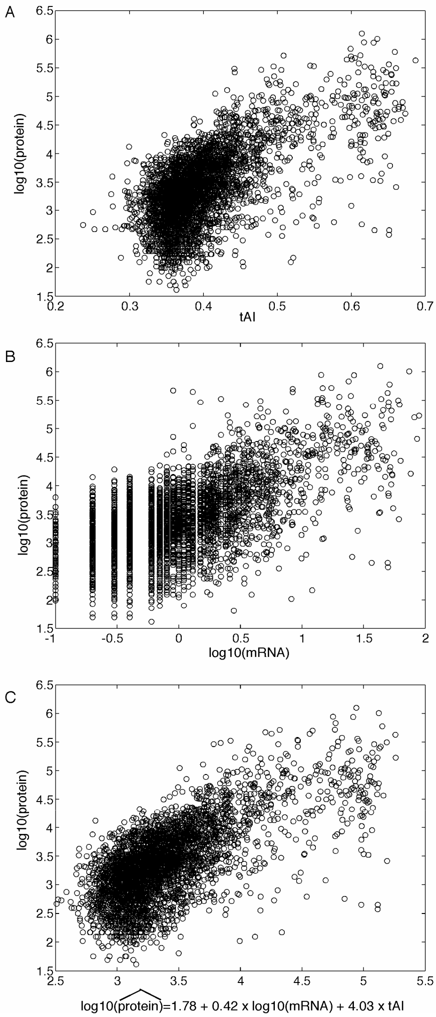

Fig. 1. The relationship of tAI and experimentally-determined mRNA levels with experimentally-determined protein levels in *S. cerevisiae*. A. log-transformed protein levels vs. tAI; B. log-transformed protein levels vs. log-transformed mRNA levels. C. experimentally-determined log-transformed protein levels vs. predicted log-transformed protein levels, obtained from multiple linear regression using both tAI and mRNA levels. mRNA and protein data were obtained from [13] and [9], respectively.

protein levels when mRNA levels are also available we computed a multiple linear regression model utilizing both tAI and log-transformed mRNA levels [13] to predict the log-transformed protein levels [9] (Fig. 1C). The model's improvement over the individual predictors seems quite modest, with the Pearson correlation coefficient of the fitted values with the log-transformed protein levels being 0.67 ($p<1e-363$). However, computation of the partial correlations indicates that each of the individual variables makes a significant contribution: the partial correlation of tAI with the log-transformed protein levels, given the log-transformed mRNA levels is R=0.34 (p=3.44e-100); and a partial correlation of R=0.29 (p=6.4e-74) for the log-transformed mRNA and protein levels, given the tAI.

Examination of the scatter-plot of tAI vs. the log-transformed protein levels, reveals that although the correlation between these two variables is highly significant, similar to the observation for mRNA levels, tAI is not a good predictor of protein levels. A possible explanation for the inaccuracies in the predictions of the tAI is that, whereas protein and transcript levels vary across different conditions, the tAI is independent of conditions. Therefore, it is possible that tAI is an indicator of the maximal potential protein levels, rather than the protein levels at a specific condition. To examine the validity of this hypothesis we capitalized on the many microarray experiments of recent years, as compared to few proteomic studies. The fact that high mRNA levels generally correspond to high protein levels, allows us to make a comparison of tAI with mRNA levels rather than protein levels. We examined 24 outliers for the correlation between the tAI and mRNA levels, that exhibited relatively high tAI (greater than 0.5), but lower levels of transcript than would be expected, and looked for experiments where these outliers were induced, using a wild-type yeast strain. This analysis was obviously limited to the conditions covered by experiments published to date, and therefore does not necessarily cover all the conditions yeast cells may experience. However, despite this limitation, in the majority of cases (21/24 cases) we could find a condition under which the ORF was at least twofold induced, with the lowest maximal induction being 3.4-fold for YLR461W, a member of the seripauperin family, during the unfolded protein response [15]. In many cases we could find an experiment for which the product of the mRNA levels at log-phase [13] and the fold-induction value was in line with the expected mRNA level. For example, YPL240C, a cytoplasmic chaperone of the HSP90 family with a tAI value of 0.60, but very modest transcript levels under log-phase growth [13], is induced 11.7-fold during a heat shock experiment from 21°C to 37°C [16]. Thus, available data support our hypothesis of tAI as an indicator of maximal protein levels under all possible conditions encountered by the cell.

2.2 tAI as a Predictor for Translational Selection in a Genome

The application of the tAI to the sequences of a genome is useful only if translational selection has played a significant part in shaping the codon usage of a genome. Thus, before selecting species for a multi-species analysis we checked whether translational selection can be detected in their genome. The effective number of codons (Nc) is a measure of the departure of codon usage in a sequence from random usage of synonymous codons, and is related to the amount of entropy in codon usage [17]. Nc reaches its maximal value (61) when codon usage is completely random, and its minimal value when only one codon is used per amino acid. Therefore, if translational selection were the only force shaping codon usage, sequences selected for optimal translation would be detected by their low values of Nc. However, codon usage, and with it Nc, is largely affected by the silent GC content (Xg), *i.e.* the percentage of codons that have guanine or cytosine at their third nucleotide position. dos Reis et al. [1] have suggested testing for the presence of translational selection in a genome, by assessing the correlation between the tAI, and the difference between f(Xg), a function predicting Nc based solely on Xg, and Nc. A strong positive correlation in this test would indicate co-adaptation between codon usage and tRNA gene copy numbers. We applied this test to eight ascomycotic yeast species, and found translational selection to be present in all of them (Table 1; Fig. 2). In spite of this, it may be that while translational selection shaped the residual codon bias in coding sequences left after accounting for the effect of silent GC, mutation pressure has been so strong that the effect of translational selection on the overall codon bias in the sequence might be minute. In such a case, changes in protein levels may be achieved in different ways, for example by raising the levels of transcript. This suggests that the test by dos Reis et al is not appropriate for testing the extent of the effect of translational selection in shaping codon usage, and as a consequence on expression. Therefore, to test the contribution of translational selection to overall codon usage, we tested the correlation between tAI and Nc. This time we expect strong negative correlation if codon usage is highly adapted to the cellular tRNA pools. We found that for seven of the species this correlation was highly significant (Table 1; Fig. 3A). However, for *A. gossypii* the magnitude of correlation was low and insignificant (Table 1; Fig. 3B), suggesting that for this species tAI would not be a good predictor of expression levels. We therefore excluded *A. gossypii* from the subsequent analysis.

Table 1. Pearson correlation of tAI with $f_l(X_g)$-Nc and with Nc in various ascomycotic yeast species

species	correlation of tAI with $f_l(X_g)$-Nc	significance	correlation of tAI with Nc	significance
A. gossypii	0.60	< 0.001	-0.38	0.384
C. albicans	0.62	0.002	-0.65	0.005
C. glabrata	0.86	<0.001	-0.79	<0.001
D. hansenii	0.78	<0.001	-0.75	<0.001
S. bayanus	0.81	<0.001	-0.73	<0.001
S. cerevisiae	0.81	<0.001	-0.79	<0.001
S. pombe	0.83	<0.001	-0.66	<0.001
Y. lipolytica	0.82	<0.001	-0.84	<0.001

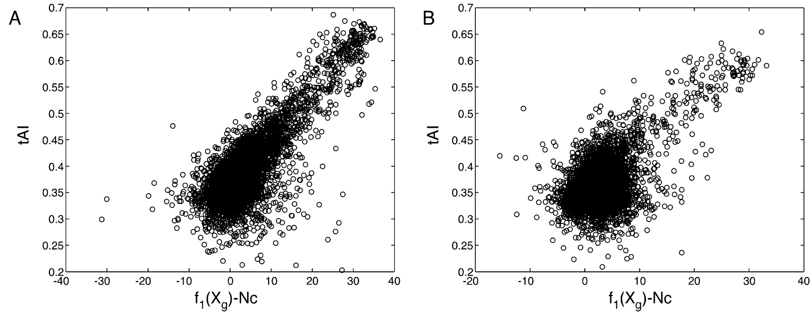

Fig. 2. tAI vs. $f_l(X_g)$-Nc for *S. cerevisiae* (A) and *A. gossypii* (B) ORFs

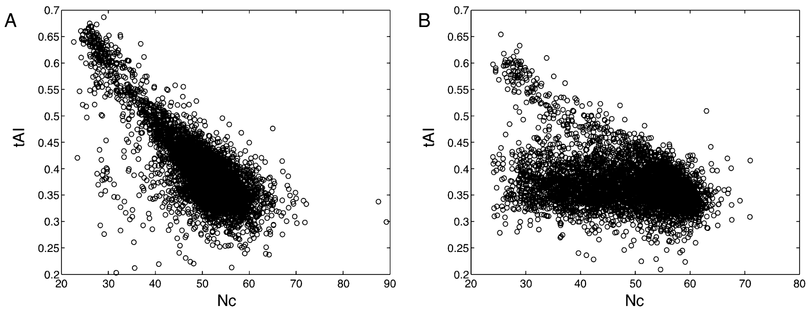

Fig. 3. tAI vs. Nc for *S. cerevisiae* (A) and *A. gossypii* (B) ORFs

In general, we can conclude that the tAI cannot be used as a predictor of protein expression levels in all genomes. It has already been suggested that translational selection may not be operating in all genomes [1]. However, our results indicate that even in cases where translational selection can be shown to be present, as indicated by the difference between f(Xg) and Nc, it may still be that translational selection makes an insignificant contribution to the overall codon bias, making the tAI an inappropriate predictor of protein expression levels.

2.3 Multi-species tAI Profiles Contain Functional Information

If phenotypic divergence between species were the product of different expression regimes, it is expected that the levels of the proteins underlying the phenotype would vary across species in a coordinated manner under the relevant conditions. In this respect the prediction of protein levels from coding sequences seems problematic, since these predictions are condition-independent. Yet, recent studies [5, 6], using CAI as a predictor for protein expression levels, showed that the profiles of predicted expression levels across species tended to be correlated for functionally interacting protein pairs, indicating that functional inferences based on predicted expression levels might be possible.

To validate the use of tAI for functional inferences we checked the behavior of tAI profiles of genes across species for a set of experimentally-determined interacting protein pairs taken from the data of [18]. We generated, using sequence similarity measures, a table of close to 5000 orthologous groups in the seven ascomycotic yeast species that showed a significant influence of translational selection over their codon usage patterns (Table 1). Over 2500 of these groups were present in at least six of the eight species, including *S. cerevisiae* and *S. pombe*, and were used for further analysis. For each orthologous group we generated a profile of predicted expression levels across species, using the tAI. We thus obtained a matrix where each column corresponds to a yeast species, and each row to an orthologous group. We compared the distribution of Pearson correlation coefficients found among the tAI profiles of protein pairs known to interact, with those of two sets of non-interacting protein pairs. The first set of non-interacting pairs, C, was obtained by calculating all possible pairs using the proteins in the interacting pairs set, and then subtracting those pairs that are known to interact. The second control, C', was a random sample of 1311 pairs from C, the same number of pairs as in the set of known physically interacting pairs. We found the correlations among the physically interacting pairs to be significantly higher

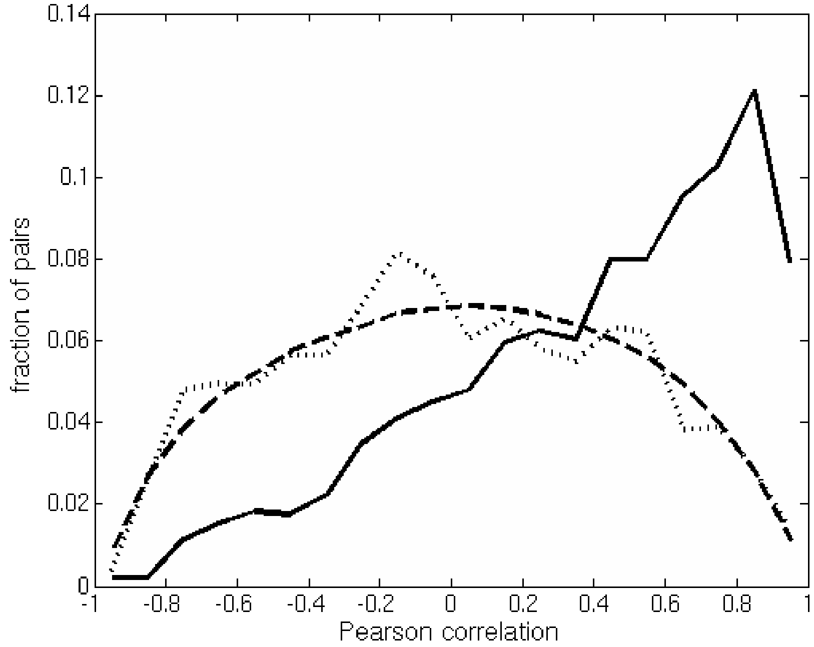

Fig. 4. Distribution of Pearson correlation coefficients among the tAI profiles of protein pairs known to physically interact (solid line), as compared to the corresponding distributions of the two sets of non-interacting proteins C (dashed line) and C' (dotted line)

than those found in both sets of non-interacting pairs (p<1e-100 for both comparisons using a Wilcoxon-Cox rank sum test; Fig. 4).

3 Conclusion

The tAI, an intuitive measure of the optimality of a coding sequence in terms of translation, correlates well with experimentally-determined protein and mRNA levels, and may be a good predictor of the maximal levels of protein under all conditions encountered by a species. The putative levels of protein predicted by the tAI tend to vary in a coordinated manner across species for physically interacting pairs, indicating the potential of this index to serve in functional inferences regarding phenotypic differences among closely-related species. However, care should be taken to apply the tAI only to genomes where translational selection can be shown to be a major force shaping codon usage in sequences.

4 Data and Methods

4.1 Species Analyzed

The yeast species used in this study are *Saccharomyces cerevisiae*, *Saccharomyces bayanus*, *Candida glabrata*, *Ashbya gossypii*, *Debaryomyces hansenii*, *Candida albicans*, *Yarrowia lipolytica* and *Schizosaccharomyces pombe*.

4.2 Protein and Coding Sequences

C. albicans protein and coding sequences were downloaded from [19]. *S. cerevisiae* and *S. bayanus* protein and coding sequences were downloaded from [20]. For *S. bayanus* several sequences may correspond to different fragments of the same ORF. We used the annotation given by [21] to merge such fragments.

For the remaining five species files in both fasta and UniProt formats were downloaded from [22]. The UniProt format files were used to construct a dictionary, linking accessions referring to nucleotide sequences to their corresponding proteins. We then downloaded, from [23] all entries corresponding to the species in question and containing a "CDS" feature, in EMBL format. A perl script utilizing BioPerl [24] was then used to go over the EMBL format file to extract coding sequences of accessions corresponding, according to the dictionary, to sequences in the protein fasta file. Coding sequences were used only if their length was at least three times the length of the protein sequence. If the coding sequence was longer than this length we assumed that this was due to an alternative initiation of the sequence, and used the last N nucleotides, where N is the expected length for the coding sequence.

4.3 tRNA Gene Copy Numbers

For all species except *C. albicans* and *S. bayanus* the tRNA gene copy numbers were obtained by applying the tRNAscan-SE software version 1.1 [25] to chromosome sequences obtained from GenBank (http://www.ncbi.nlm.nih.gov/Genbank). For *S. bayanus* we used the tRNA gene copy numbers of the closely related *S. cerevisiae*. Although a list of the tRNA genes for this species is available [21], the low total number of protein-coding genes available for it and the other two *sensu stricto* species sequenced in the same project (less than 5000 in each of the three species, compared to close to 6000 in *S. cerevisiae*), indicates that the quality of the genome sequence may not be high enough to reliably determine the copy numbers of tRNA genes. The strong conservation of synteny between *S. bayanus* and *S. cerevisiae* [21] and the relatively short time that has passed since their divergence (~20 million years ago) makes the use of the tRNA gene copy numbers from *S. cerevisiae* a conservative choice.

For *C. albicans* we extracted the tRNA gene counts from [19].

4.4 Calculation of tAI, Nc, $f_1(X_g)$-Nc, Their Correlation and Its Significance for Coding Sequences

The tAI method is described in detail in [1]. Briefly, the method entails calculating a weight for each of the sense codons, derived from the copy numbers of all the tRNA types that recognize it. For a given coding sequence, the tAI value is then the geometric mean of the weights of all its sense codons (stop codons were ignored when encountered). To calculate the tAI for coding sequences we used the codonR scripts supplied by [1], downloaded from http://people.cryst.bbk.ac.uk/~fdosr01/tAI/, which we modified to include the first codon, as well as other methionines. The effective number of codons (Nc,[17]) was calculated with the modified version of the codonW program, supplied by [1], which was downloaded from the same site. This version of codonW was further modified to accommodate the alternative yeast nuclear code used by *D. hansenii* [26]and *C. albicans* [27]. The significance of the observed correlation of Nc with tAI was calculated by permuting the tAI weights of the sense codons 1000 times. Each such permutation was then used to compute the correlation of Nc with the tAI calculated using the randomized weights. These calculations were done using scripts downloaded from the same site and the R software for statistical computing (http://www.r-project.org).

4.5 Generation of a Table of Orthologous Groups

The orthologous groups were constructed using a *S. cerevisiae*-centered methodology. We constructed, using the inparanoid algorithm [28], six lists of orthologous groups containing genes from only two species – *S. cerevisiae* and one of the other six species. These two-way groups were then merged to obtain orthologous groups potentially encompassing all species in the sample.

The generation of the lists of two-species orthologous groups utilized the inparanoid algorithm [28]without an outgroup and without bootstrapping. We kept only sequences that were assigned a confidence value of at least 25% for their membership in the group. There is a discrepancy between the inparanoid algorithm as reported in [28], and the program supplied by the authors at http://inparanoid.cgb.ki.se/: while the paper specifies that the matched segment between two sequences must cover at least 50% of the longer sequence in order for the sequences to be considered homologous, the program applies this cutoff to the shorter sequence. In order to avoid domain-level matches, we modified the inparanoid program to reflect the algorithm as presented in the paper.

The second part, the merger of the six two-species lists into one seven-species list, was achieved by iteratively adding the two-species lists. The order of processing of the lists was dictated by the relative closeness of the second species in the list to *S. cerevisiae* (*S. cerevisiae* was included in all lists): starting with *S. cerevisiae*'s most distant relative (*S. pombe*) and ending with its closest relative (*S. bayanus*). *C. albicans* and *D. hansenii* are equidistant from *S. cerevisiae*, and we arbitrarily chose to first analyze the *C. albicans* list. Each iteration consisted of going over all orthologous groups in the list being processed. For each such group, if its *S. cerevisiae* genes were found in a group obtained in a previous iteration, the two groups were merged; otherwise, if the genes in the group appeared after the divergence of *S. cerevisiae* from the previously analyzed species, a new group was created. Note that if a duplication had occurred after the divergence from the previously analyzed species then more than one group may be merged with the same pre-existing group. The order of merger, *i.e.* according to the order of divergence from *S. cerevisiae*, ensures that there will be no ambiguity as to which pre-existing group to add a currently analyzed group.

4.6 Generation of a Matrix of Predicted Expression Levels Across Species

We combined the orthologous groups table with tAI scores to create a matrix of predicted expression levels across species. In cases where the orthologous group contained several paralogs we used the maximal tAI score among them. We discarded all profiles that had no representative from *S. pombe*. The resultant matrix of predicted expression was then submitted for preprocessing at the GEPAS server [29]: profiles with more than 30% missing values were removed, missing values in the remaining profiles were imputed using the KNNimpute algorithm with k=15. This left us with 2592 profiles. Each column of the matrix was subsequently normalized, so that for each species the mean and standard deviation would be 0 and 1, respectively.

4.7 Analysis of Outliers Having High tAI, But Relatively Low mRNA Expression Levels

We analyzed two groups of outliers in the comparison of tAI and experimentally determined mRNA levels obtained from [13]: ORFs with 0.6≤tAI and transcript levels at most 10 (9 ORFs), and ORFs with 0.5≤tAI<0.6 and transcript levels at most 1 (20 ORFs). Since mRNA levels are subject to noise, results for the same condition may vary across experiments, and what may seem as an outlier using the data of one experiment may not be an outlier using the data of another experiment. We therefore filtered out from the set of outliers five ORFs that were not outliers using another dataset obtained under similar conditions (dataset GSM6711 corresponding to one of the control samples in the study of [30]). The control dataset was downloaded from the GEO database [31] (http://www.ncbi.nlm.nih.gov/geo). We then used the "expression connection" tool at the SGD database site [32] to obtain the maximal induction levels of the outlying ORFs, taking care to consider only experiments in which a wild-type strain was used.

4.8 Pairs of Physically Interacting Genes in *S. crevisiae*

We extracted 2301 pairs of proteins from the supplementary data of von Mering et al [18] by filtering out those protein pairs that were marked as "previously annotated: no". We further filtered out pairs of paralogs belonging to the same orthologous groups, and pairs where at least one member was not in the data matrix, leaving 1311 pairs of interacting proteins.

References

1. dos Reis, M., Savva, R., Wernisch, L.: Solving the riddle of codon usage preferences: a test for translational selection. Nucleic Acids Res **32** (2004) 5036-5044
2. Pellegrini, M., Marcotte, E.M., Thompson, M.J., Eisenberg, D., Yeates, T.O.: Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. Proc Natl Acad Sci U S A **96** (1999) 4285-4288
3. Avidor-Reiss, T., Maer, A.M., Koundakjian, E., Polyanovsky, A., Keil, T., Subramaniam, S., Zuker, C.S.: Decoding cilia function: defining specialized genes required for compartmentalized cilia biogenesis. Cell **117** (2004) 527-539
4. Martin, M.J., Herrero, J., Mateos, A., Dopazo, J.: Comparing bacterial genomes through conservation profiles. Genome Res **13** (2003) 991-998
5. Fraser, H.B., Hirsh, A.E., Wall, D.P., Eisen, M.B.: Coevolution of gene expression among interacting proteins. Proc Natl Acad Sci U S A **101** (2004) 9033-9038
6. Lithwick, G., Margalit, H.: Relative predicted protein levels of functionally associated proteins are conserved across organisms. Nucleic Acids Res **33** (2005) 1051-1057

7. Sharp, P.M., Li, W.H.: The codon Adaptation Index--a measure of directional synonymous codon usage bias, and its potential applications. Nucleic Acids Res **15** (1987) 1281-1295
8. Percudani, R., Pavesi, A., Ottonello, S.: Transfer RNA gene redundancy and translational selection in Saccharomyces cerevisiae. J Mol Biol **268** (1997) 322-330
9. Ghaemmaghami, S., Huh, W.K., Bower, K., Howson, R.W., Belle, A., Dephoure, N., O'Shea, E.K., Weissman, J.S.: Global analysis of protein expression in yeast. Nature **425** (2003) 737-741
10. Greenbaum, D., Jansen, R., Gerstein, M.: Analysis of mRNA expression and protein abundance data: an approach for the comparison of the enrichment of features in the cellular population of proteins and transcripts. Bioinformatics **18** (2002) 585-596
11. Lavner, Y., Kotlar, D.: Codon bias as a factor in regulating expression via translation rate in the human genome. Gene **345** (2005) 127-138
12. Friberg, M., von Rohr, P., Gonnet, G.: Limitations of codon adaptation index and other coding DNA-based features for prediction of protein expression in Saccharomyces cerevisiae. Yeast **21** (2004) 1083-1093
13. Holstege, F.C., Jennings, E.G., Wyrick, J.J., Lee, T.I., Hengartner, C.J., Green, M.R., Golub, T.R., Lander, E.S., Young, R.A.: Dissecting the regulatory circuitry of a eukaryotic genome. Cell **95** (1998) 717-728
14. Gygi, S.P., Rochon, Y., Franza, B.R., Aebersold, R.: Correlation between protein and mRNA abundance in yeast. Mol Cell Biol **19** (1999) 1720-1730
15. Travers, K.J., Patil, C.K., Wodicka, L., Lockhart, D.J., Weissman, J.S., Walter, P.: Functional and genomic analyses reveal an essential coordination between the unfolded protein response and ER-associated degradation. Cell **101** (2000) 249-258
16. Gasch, A.P., Spellman, P.T., Kao, C.M., Carmel-Harel, O., Eisen, M.B., Storz, G., Botstein, D., Brown, P.O.: Genomic expression programs in the response of yeast cells to environmental changes. Mol Biol Cell **11** (2000) 4241-4257
17. Wright, F.: The 'effective number of codons' used in a gene. Gene **87** (1990) 23-29
18. von Mering, C., Krause, R., Snel, B., Cornell, M., Oliver, S.G., Fields, S., Bork, P.: Comparative assessment of large-scale data sets of protein-protein interactions. Nature **417** (2002) 399-403
19. Arnaud, M.B., Costanzo, M.C., Skrzypek, M.S., Binkley, G., Lane, C., Miyasato, S.R., Sherlock, G.: "Candida Genome Database" http://www.candidagenome.org/. (14 August 2005)
20. Balakrishnan, R., Christie, K.R., Costanzo, M.C., Dolinski, K., Dwight, S.S., Engel, S.R., Fisk, D.G., Hirschman, J.E., Hong, E.L., Nash, R., Oughtred, R., Skrzypek, M., Theesfeld, C.L., Binkley, G., Dong, Q., Lane, C., Sethuraman, A., Weng, S., Botstein, D., Cherry, J.M.: "Saccharomyces Genome Database" ftp://ftp.yeastgenome.org/yeast/. (16 June 2005)
21. Kellis, M., Birren, B.W., Lander, E.S.: Proof and evolutionary analysis of ancient genome duplication in the yeast Saccharomyces cerevisiae. Nature **428** (2004) 617-624
22. Pruess, M., Kersey, P., Apweiler, R.: The Integr8 project--a resource for genomic and proteomic data. In Silico Biol **5** (2005) 179-185
23. Kanz, C., Aldebert, P., Althorpe, N., Baker, W., Baldwin, A., Bates, K., Browne, P., van den Broek, A., Castro, M., Cochrane, G., Duggan, K., Eberhardt, R., Faruque, N., Gamble, J., Diez, F.G., Harte, N., Kulikova, T., Lin, Q., Lombard, V., Lopez, R., Mancuso, R., McHale, M., Nardone, F., Silventoinen, V., Sobhany, S., Stoehr, P., Tuli, M.A., Tzouvara, K., Vaughan, R., Wu, D., Zhu, W., Apweiler, R.: The EMBL Nucleotide Sequence Database. Nucleic Acids Res **33** (2005) D29-33

24. Stajich, J.E., Block, D., Boulez, K., Brenner, S.E., Chervitz, S.A., Dagdigian, C., Fuellen, G., Gilbert, J.G., Korf, I., Lapp, H., Lehvaslaiho, H., Matsalla, C., Mungall, C.J., Osborne, B.I., Pocock, M.R., Schattner, P., Senger, M., Stein, L.D., Stupka, E., Wilkinson, M.D., Birney, E.: The Bioperl toolkit: Perl modules for the life sciences. Genome Res **12** (2002) 1611-1618
25. Lowe, T.M., Eddy, S.R.: tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res **25** (1997) 955-964
26. Tekaia, F., Blandin, G., Malpertuy, A., Llorente, B., Durrens, P., Toffano-Nioche, C., Ozier-Kalogeropoulos, O., Bon, E., Gaillardin, C., Aigle, M., Bolotin-Fukuhara, M., Casaregola, S., de Montigny, J., Lepingle, A., Neuveglise, C., Potier, S., Souciet, J., Wesolowski-Louvel, M., Dujon, B.: Genomic exploration of the hemiascomycetous yeasts: 3. Methods and strategies used for sequence analysis and annotation. FEBS Lett **487** (2000) 17-30
27. Sugita, T., Nakase, T.: Non-universal usage of the leucine CUG codon and the molecular phylogeny of the genus Candida. Syst Appl Microbiol **22** (1999) 79-86
28. Remm, M., Storm, C.E., Sonnhammer, E.L.: Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. J Mol Biol **314** (2001) 1041-1052
29. Herrero, J., Al-Shahrour, F., Diaz-Uriarte, R., Mateos, A., Vaquerizas, J.M., Santoyo, J., Dopazo, J.: GEPAS: A web-based resource for microarray gene expression data analysis. Nucleic Acids Res **31** (2003) 3461-3467
30. Bulik, D.A., Olczak, M., Lucero, H.A., Osmond, B.C., Robbins, P.W., Specht, C.A.: Chitin synthesis in Saccharomyces cerevisiae in response to supplementation of growth medium with glucosamine and cell wall stress. Eukaryot Cell **2** (2003) 886-900
31. Barrett, T., Suzek, T.O., Troup, D.B., Wilhite, S.E., Ngau, W.C., Ledoux, P., Rudnev, D., Lash, A.E., Fujibuchi, W., Edgar, R.: NCBI GEO: mining millions of expression profiles--database and tools. Nucleic Acids Res **33** (2005) D562-566
32. Balakrishnan, R., Christie, K.R., Costanzo, M.C., Dolinski, K., Dwight, S.S., Engel, S.R., Fisk, D.G., Hirschman, J.E., Hong, E.L., Nash, R., Oughtred, R., Skrzypek, M., Theesfeld, C.L., Binkley, G., Dong, Q., Lane, C., Sethuraman, A., Weng, S., Botstein, D., Cherry, J.M.: "Saccharomyces Genome Database" http://www.yeastgenome.org/. (1 February 2006)