מכון ויצמן למדע

WEIZMANN INSTITUTE OF SCIENCE

Thesis for the degree
Master of Science

חבור לשם קבלת התואר
מוסמך למדעים

By
Asaf Carmi

מאת
אסף כרמי

האבולוציה של הרצפים המקודדים ומאגר ה-tRNA ותפקידה בעיצוב יעילות התרגום

The evolution of coding sequences and tRNA pools, and its role in shaping translation efficiency

Advisor
Prof. Yitzhak Pilpel

מנחה
פרופ' יצחק פלפל

December 2008

כסלו תשס"ט

Submitted to the Scientific Council of the
Weizmann Institute of Science
Rehovot, Israel

מוגש למועצה המדעית של
מכון ויצמן למדע
רחובות, ישראל

Acknowledgements

I would like to thank a number of people, who have helped me, in numerous ways, during the progress of my Masters thesis. First and foremost, I would like to thank my advisor, Prof. Tzachi Pilpel, for his support, encouragement and helpful discussions and for giving me the opportunity to spend two exciting years in his lab. I would also like to thank Dr. Itay Furman, who was always available for discussions, support and suggestions. Much of what I have learned during this past two years I have learned from him.

I would like to thank my collaborator in the viral tRNA project, Keren Limor-Waisberg. Her motivation and enthusiasm have inspired me, and our discussions, both scientific and friendly, were helpful and enjoyable. I would also like to thank members of the Pilpel lab, who were always willing to discuss my work, help, and give moral support. I would especially like to thank Hila Gingold for our fruitful discussions and for her willingness to assist in any need.

Finally, I would like to thank my family and friends, for encouraging me and providing support and a listening ear whenever needed.

Table of contents

Abstract

Translation is a highly regulated process in cells. An important means to control the efficiency of translation is by tuning the adaptation of genes' codon usage to the cellular pool of tRNAs. The aim of my study was to explore the co-evolution between coding sequences and tRNA pools across multiple organisms and to establish its role in shaping translation efficiency. The study consists of three parts.

In the first part, I used a novel computational approach based on the well-known tRNA adaptation index (tAI), to quantitatively break down the differences in translation efficiency between orthologous genes, into contributions from differences in codon usage ("*cis*" changes), and differences in tRNA pools ("*trans*" changes). Using eight fully sequenced yeast species, I have found that changes in the coding sequence make the most significant contribution to the evolution of translation efficiency. This framework also provided quantitative means for detecting co-evolution of gene sequences and tRNA pools.

In the second part I used the tAI as a way to probe the process of viral infection. I chose *cyanophage* Syn9 as a test case, since its genome encodes for six tRNA genes, and it was found to infect a wide range of *cyanobacteria* hosts. I computed the translation efficiency of the viral genes in the background of several hosts with or without its own tRNA pool. Based on the differences found, I have concluded that the viral tRNA genes have evolved to boost the translation efficiency of its own genes in specific hosts. Specifically, the virion structural proteins benefit the most from the addition of the viral encoded tRNAs. Similar computation on the genes of the host that provides the most favorable background revealed that among its annotated genes, cell envelope and transport proteins would benefit the most from the addition of viral tRNAs. Interestingly, uncharacterized proteins constitute 70% of the host genes which benefited the most from the viral tRNA pool.

In the last part I investigated the spatial patterns in the translation efficiency. I observed a pattern of gradual increase in the translation efficiency along coding sequences, and that this pattern is conserved among various yeast species and in *E. coli*. I have also shown that the coding sequences and the tRNA pool have co-evolved in order to conserve this pattern. Computer simulations that I have designed and executed suggest that this pattern reduces the number of ribosomal collisions on

mRNA transcripts, therefore potentially increasing the efficiency of the translation process.

1. Introduction

The translation process is a highly regulated process. Yet on a genome-wide level it is much less studied compared to transcription. Regulation of the translation process is occurring in three major stages: initiation, elongation and termination. Unlike initiation and termination, the machinery used during translation elongation has been highly conserved across the three kingdoms of life. Therefore, the mechanisms underlying the elongation process are assumed to be the same in eukaryotes, bacteria and archea (Kapp and Lorsch 2004). In this work I use computational tools to examine the efficiency of the elongation process in various contexts. In the first part I will show a comparative study in yeast demonstrating the evolution of this regulation layer. In the second part, which was done with collaboration with Keren Limor Waisberg and Prof. Avigdor Scherz, I will show how viruses control the efficiency of translation of particular gene sets to enhance their infection capability. In the third part I will study the local effects of the elongation efficiency. I will start by defining more accurately what "efficiency" means in the context of the elongation process.

1.1. Translation efficiency

The efficiency of the elongation process (termed translation efficiency, elongation efficiency or efficiency throughout this work) is determined by many factors, such as mRNA secondary structure (Gray and Hentze 1994) and ribosomal electrostatic charges (Lu and Deutsch 2008). The common way to measure translation efficiency of a coding sequence is by the extent of the adaptation of its codon usage to the tRNA cellular pools (Sharp and Li 1987), (dos Reis et al. 2004) which serves as a surrogate measure for its speed of translation. This definition stems from an early observation of a trend of increasing codon usage bias with increasing gene expression levels in a sample of *E. coli* genes (Sharp and Li 1986), and that tRNA concentrations are rate limiting in the elongation of nascent peptides (Varenne et al. 1984).

The translation efficiency, as defined above, has also been shown to be correlated with translation rate and accuracy (Akashi 2003), phenotypic divergence of yeast species (Man and Pilpel 2007) and to also play part in protein functionality (Kimchi-Sarfaty et al. 2007).

Many measures have been proposed to evaluate the translation efficiency. Those are roughly divided into two categories. In the first category are measures which test for deviation of the codon usage of genes from the equal use of synonymous codons. From those, the most used is the effective number of codons (Nc) (Wright 1990). The second category includes measures that test the conformance of a sequence's codon usage to a 'translationally optimal' codon usage. Translationally optimal codons are codons which correspond to abundant tRNAs in the cellular tRNA pool. However, experimental data about the concentrations of the various tRNA types in the cell are available for very few species and under a limited set of growth conditions (Ikemura 1982),(Kanaya et al. 1999).

The Codon Adaptation Index (CAI) (Sharp and Li 1987) which belongs to the second category, is the most widely used measure of translation efficiency. It addresses the lack of data about tRNA concentration by using a reference set of genes that are known to be highly expressed, and hence assumed to have optimal codon usage in terms of translation. All remaining genes belonging to the same genome are then scored according to the similarity of their codon usage to that of the reference set. The disadvantage of this solution is that in order to calculate the CAI of an organism's genes, knowledge of a set of highly expressed genes in this organism is necessary and this data is not widely available. Furthermore, since the measure is biased towards the highly-expressed genes, the mapping between codon usage dissimilarity and translation efficiency is questionable. In addition, while the codons represent the "demand" in the process, the tRNAs represent the "supply", a layer that the CAI does not treat explicitly.

The tRNA Adaptation Index, tAI (dos Reis et al. 2004), a relatively recent index of translation efficiency, uses the tRNA genes copy numbers (tGCNs) in the genome as a means to calculate the translation efficiency, by assigning weights to each codon based on abundance of its cognate tRNA taking into account wobble interactions. Using the tGCN as a surrogate measure for the cellular abundances of tRNAs is justified by several observations. First, it has been observed that the *in vivo* concentration of a tRNA bearing a certain anticodon is highly proportional (r=0.91 for *S. cerevisiae*) to the number of gene copies coding for this tRNA type (Percudani et al. 1997),(Kanaya et al. 1999). Second, a recent study showed that in *S. cerevisiae* the promoters of many of the tRNA genes have a low predicted affinity to the

nucleosome, suggesting a constitutive expression with little transcriptional regulation capacity (Segal et al. 2006). Thus, for fully sequenced genomes, the relative concentrations of the various tRNAs in the cell, and therefore the optimality of the various codons in terms of translation, can be approximated using the respective tRNA gene copy numbers in the genome.

The tAI has been shown to be highly correlated (r=0.63 for *S. cerevisiae*) to protein expression levels. It was found that even among genes with similar transcript levels, higher tAI often corresponds to higher protein abundance (Man and Pilpel 2007).

1.2. Breaking down the evolution of genes expression to *Cis* and *Trans* components

The evolution of gene expression was suggested by many to serve as a major driving force in the evolution and divergence of species (Dennis A. Powers 1998), (Prud'homme, Gompel et al. 2007), (Ihmels, Bergmann et al. 2005). For instance it is often claimed that most of the differences between human and chimpanzee are not at the gene content level, but rather in the ways orthologous genes are differentially regulated in each species (Gilad, Oshlack et al. 2006). When analyzing orthologous genes that are differentially regulated across species, researchers often differentiate between two types of changes. The first, are changes that occur within the genes, or in near-by regulatory regions, usually referenced as changes in "*Cis*". The second are changes that occur elsewhere in the cell, presumably among the regulators of the genes, usually referenced as "*Trans*" changes. The distinction between *cis* and *trans* changes in driving overall expression differences of orthologous genes across various species is of significant importance, since they act on a different scale. While *cis* changes usually affect expression on a gene-specific level, *trans* changes have the potential to affect expression of entire gene modules and networks.

Wittkop et al. (Wittkopp, Haerum et al. 2004) introduced a system to detect the contribution of *cis* and *trans* modification to the difference in expression of orthologous genes in *Drosophila*, and found that most of the differences can be explained by a *cis* factor modification. Studies performed in yeast (Wang, Sung et al. 2007) concluded that *trans* factors account to most of the differences observed in the transcription program of orthologous genes in multiple yeast species.

Yet, so far the study of the effect of *cis* and *trans* changes in the evolution of gene expression was mostly confound to the level of transcription. Moving into the translation realm, those concepts can be applied to changes in the translation efficiency of genes. Since the translation efficiency is measured by the extent of the adaptation of a gene's codon usage to the tRNA cellular pools, differences in the translation efficiency can arise due to changes in the tRNA pool or changes in the coding sequence. Changes to the tRNA pool of an organism can be viewed as a *trans* factor affecting the efficiency, since their effect is not limited to a single gene, but rather to entire gene sets, and even potentially to the entire genome. Changes in the coding sequence of individual genes, which affect their adaptation to the tRNA pool, can be considered as a *cis* factor.

In the transcription realm there is no simple computational means to predict and assess the efficiency at which a gene is transcribed. Thus, we have to turn to lab experiments to measure the efficiency and the causes for changes in efficiency. However, the tAI serves as simple computational means to gauge for translation efficiency, and can thus provide us with a good computational measure to test the causes for differences in the translation efficiency between orthologous genes. This tool is suitable for this need since it incorporates both the tRNA availability and the coding sequence adaptation to it.

In the first part of this thesis, I developed a computational method to measure how the coding sequence and the tRNA pool co-evolved to create the translation efficiency differences as depicted by the tAI measure. I then applied the method to eight fully sequenced yeast species in order to study how those changes create differences in the translation efficiency of genes. I have found that changes in the translation efficiency occurred through both *cis* and *trans* changes, and through different interplays between the two factors.

1.3. The effect of viral tRNA genes on the translation efficiency of viruses and their hosts

In the second part, I chose to focus on a special case of *trans* changes – the effect of viral tRNA genes on the translation efficiency of viral and host genes. The presence of tRNA genes carried by phages was first discovered in the T4 phage (Weiss et al.

1968). The availability of fully sequenced viral genomes revealed many phages which carry tRNA genes in their genome (Bailly-Bechet et al. 2007).

The reasons for the presence of tRNA genes in phages are still enigmatic. Several assumption have been made, including allowing the phage to be resistant to anticodon nucleases in the host (Kaufmann 2000), (Blanga-Kanfi et al. 2006), allowing a better integration of lysogenic phages in the host chromosome (Carlos Canchaya 2004), (Yinling Tan 2007), and increasing the translation efficiency of the viral genes (Kunisawa 1992), (Kunisawa 2000), (Bailly-Bechet et al. 2007).

Studies on the T4 bacteriophage revealed a correlation between the virus structural proteins codon usage and the tRNA genes carried by the virus. They also revealed that lowly expressed virion protein are more adapted to use the viral tRNAs, while highly expressed proteins are more adapted to the host tRNAs (Kunisawa 1992). Deletion of the tRNA genes from the phage resulted in lower burst sizes (Wilson 1973), indicating a significant role for the viral tRNAs in expressing the viral proteins. A recent study on several viral and host genomes showed high correlation between viral codon usage and viral tRNA genes, mostly in virulent phages (Bailly-Bechet et al. 2007).

While studies on the functional role of tRNA genes are available for several viruses, they were all performed on a limited set of genes, and estimated the adaptation of the genes to the viral tRNA pool based on the codon usage. The tAI, by taking into account the tRNA gene copy numbers, can directly measure the contribution of the viral tRNA genes to the translation efficiency on a whole genome scale. An interesting point that was overlooked in past studies is the effect of the viral tRNA genes on the expression of the host genome.

I have shown in this study, using the *cyanophage* Syn9 as a test case, that viral tRNA genes have been highly optimized to work under a specific host's genetic background, and that these tRNA molecules are adapted to improve specific functional aspects, both in the virus and in its host.

1.4. Spatial patterns in translation efficiency

In the last part of my study, I chose to look at the translation efficiency not as a global property of a gene, but rather as a local property of the sequence order. Changes in the codon usage preferences along the transcript have been reported for many organisms,

both prokaryotes (Bulmer 1988), (Ohno et al. 2001) and eukaryotes (Kliman and Eyre-Walker 1998), (Drummond and Wilke 2008). Such changes were often linked to translation efficiency, and were suggested to affect the translation rate (Liljenstrom and Vonheijne 1987) in order to assist in the regulation of gene expression (Chen and Inouye 1994), the translation accuracy (Drummond and Wilke 2008), and the protein folding (Widmann, Clairo et al. 2008). However, other selection forces may play a role in shaping the local codon bias, such as regulation of mRNA secondary structure (Eyre-Walker and Bulmer 1993) and alternative splicing (Kliman and Eyre-Walker 1998).

In *E. coli*, perhaps the most studied organism in this field, it was shown that codon usage bias near the start codon and near the stop codon of a gene is weaker than in the middle (Bulmer 1988). It was suggested that this might be selected for to prevent the formation of mRNA secondary structures that might interfere with ribosome binding (Eyre-Walker and Bulmer 1993), (Adam 1996).

Chen and Inouye (Chen and Inouye 1990) showed that rarely used codons in *E. coli* are used preferentially within the first 25 codons. This was later shown to be true in many other bacteria (Ohno et al. 2001). Removing rare codons from the region close to the initiation site of a gene dramatically increased its expression (Chen and Inouye 1994), (Vervoort et al. 2000).

Computer simulations of ribosomal movement on the mRNA showed that the location of rarely used codons on the mRNA can affect the translation time, number of ribosomes on the transcript and the time to reach steady state in the translation process (Zhang et al. 1994), (Mitarai et al. 2008). It has also been shown that clusters of low usage codons appear in a wide variety of genes and organisms (Zhang et al. 1994).

The common translation efficiency methods such as the Nc, CAI and tAI (Wright 1990), (Sharp and Li 1987), (dos Reis et al. 2004) give each gene a single score based on a weighted averaged translation efficiency of each of its codons. These scoring methods, although very informative, fail to capture the spatial patterns in the translation efficiency. The tAI, due to its codon based scoring, can be easily adapted to create a gene's translation efficiency profile.

I have shown in this study the existence of spatial patterns in translation efficiency that correspond to the previously reported codon bias patterns, and that these patterns are conserved in a wide variety of organisms. Specifically I found that the beginnings

of genes are often translated slower than their middle sections, while their ends are translated faster. I have also shown that both the tRNA pools and the coding sequences are required to co-evolve in order to conserve this pattern. Looking for a potential function of this conserved design, I have shown that the observed patterns can be used to reduce the number of ribosome collisions on the mRNA, a possibility that was not largely discussed in the literature.

2. Results

2.1. *Cis* and *Trans* changes affecting translation efficiency

In the context of transcription, researches often distinguish between several mechanisms that cause mRNA expression differences between genes. The differences can be due to "*cis*" changes, which are usually changes in the promoter or other regulatory regions of a gene, "*trans*" changes, which are usually changes in the expression and/or activity of the upstream regulators, or both. Distinguishing between those mechanisms is interesting, since *cis* elements affect only the expression of their associated gene, while *trans* elements have the potential to coordinate an effect on large sets of genes. Here I extend those concepts to the translation context, to study differences in the translation efficiency of orthologous genes across species. *Cis* changes can be viewed as changes in the coding sequence of a gene, while *trans* changes can be viewed as changes to the cellular tRNA pool of an organism.

Given two orthologous genes in two species with different translation efficiencies, one can ask whether the differences arose from the different tRNA pools in those organisms (i.e. *trans* effect), differences in their coding sequence (i.e. *cis* effect) or due to a combination of the two factors. Figure 1 illustrates the available causes to the difference in translation efficiency between two orthologs.

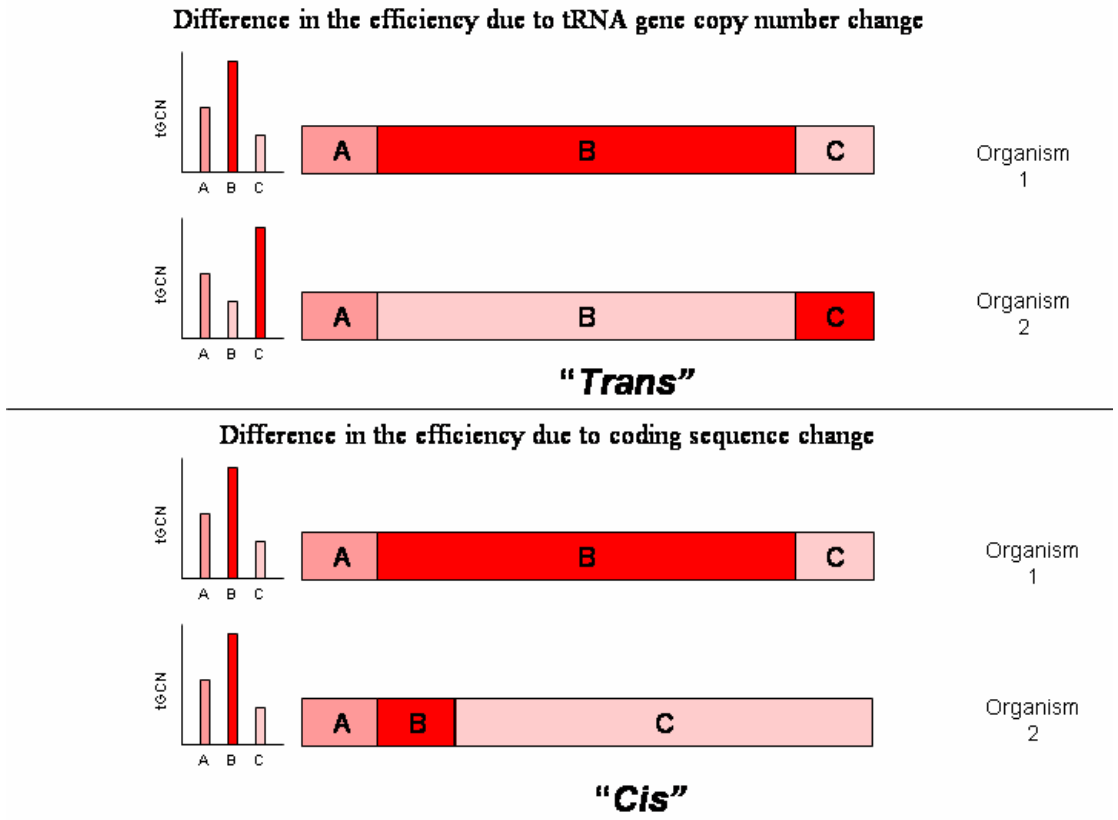


Figure 1: "*Cis*" and "*trans*" changes in the translation efficiency. The figure describes a pair of orthologous genes in two organisms, 1 and 2, in which the gene in organism 1 has higher translation efficiency. In each panel, two ortholog genes with 3 types of codons (represented as A, B, C) are described. The bar plot on the left, and the color intensity, reflect the tRNA copy numbers for codons A, B, and C. The horizontal bars represent the codon usage; segment size depicts the codon frequencies. In both panels, the translation efficiency of organism 1's gene is higher than Organism 2, since gene 2 uses more codons with corresponding low tGCN tRNAs. Two alternative evolutionary mechanisms could account for the difference in the translation efficiency: In the top panel, while codon frequency is the same in the two species, the difference in the translation efficiency is due to a shift in the tRNA gene copy numbers between codons B and C, hence a "*trans*" change. In the bottom panel, the difference in the translation efficiency is due to a change in the codon usage frequency of codons B and C, while the tRNA pool has not changes, this is hence denoted a change in "*cis*".

In this part I will develop a computational method to evaluate these elements using the tAI, and apply it to several yeast species.

2.1.1. Decomposition of the tAI change to its components

The tAI of a gene is the geometric mean of the tAI of each of its codon. i.e.:

$$tAI_g = \left(\prod_{k=1}^{\ell_g} tAI^{i^k} \right)^{1/\ell_g}$$

where tAI^{i^k} is the tAI of codon type *i*, in position *k* (see Methods and dos Reis et al. 2004). Note that the order in which the codons appear in the sequence does not affect the tAI. Therefore, this formula can be rewritten as

$$tAI_g = \prod_{i=1}^{61} tAI^{i\ n(i)/\ell_g} = \prod_{i=1}^{61} tAI^{i\ f^i}$$

where *n(i)* is the number of times codon i appears in the gene and f^i is the fraction of codon i in the gene.

Looking at a pair of orthologous genes in two organisms, A and B, one can calculate their tAI ratio:

$$\text{diff}(tAI_A, tAI_B) = \frac{tAI_A}{tAI_B} = \frac{\prod_{i=1}^{61} tAI_A^{i\ f^{i_A}}}{\prod_{i=1}^{61} tAI_B^{i\ f^{i_B}}} \quad \textbf{(Equation 1)}.$$

where tAI_A^i and tAI_B^i are the tAI of codon i in organism A and B respectively, and f_A^i and f_B^i are the fractions of the codon in the gene that comes from species A and B respectively.

For each codon *i*, and a pair of orthologous genes in two species, A and B, one can define the "common codon usage" as:

$$\begin{aligned} f^i_{common} &= \min(f^i_A, f^i_B) \\ f^i_{uniqueA} &= f^i_A - f^i_{common} \quad \textbf{(Definition 1)} \\ f^i_{uniqueB} &= f^i_B - f^i_{common} \end{aligned}$$

Defining the common and unique part enables the decomposition of the tAI, and consequently the tAI difference into two parts:

$$tAI_A = \prod_i tAI_A^{i\ f^i_{common}} \times tAI_A^{i\ f^i_{uniqueA}},\ tAI_B = \prod_i tAI_B^{i\ f^i_{common}} \times tAI_B^{i\ f^i_{uniqueB}}$$

and

$$\text{diff}(tAI_A, tAI_B) = \frac{\prod_i tAI_A^{i\ f^i_{common}}}{\prod_i tAI_B^{i\ f^i_{common}}} \times \frac{\prod_i tAI_A^{i\ f^i_{uniqueA}}}{\prod_i tAI_B^{i\ f^i_{uniqueB}}} \textbf{(Equation 2)}.$$

This decomposition allows eliminating the *cis* component from the left side of the multiplication. The left side of the multiplication involves the same codon usage

frequency, such that any differences contributed by this part are purely due to differences in the tRNA gene copy numbers (tGCNs) of the two organisms, and thus are "*trans*" changes by definition. Note that the tAI of a codon takes into account the wobble interactions, and thus is not a direct measure for the tGCN. However, since the wobble coefficients are constant for every organism, differences in the tAI reflect differences in the tGCN.

Differences in the right part can be either due to "*cis*", if codons are sharing the same tGCN, or "co-evolution", if codons have differences in their tGCN. To differentiate between those components, in the same manner as in definition 1, one need to define the common and unique tGCNs *organisms* are sharing. As explained above, this can be done using the tAI values as they reflect the tGCNs. Thus, one can define the common and unique tAIs:

$$tAI^{i}_{common} = \min(tAI^{i}_{A}, tAI^{i}_{B})$$
$$tAI^{i}_{uniqueA} = \frac{tAI^{i}_{A}}{tAI^{i}_{common}}, tAI^{i}_{uniqueB} = \frac{tAI^{i}_{B}}{tAI^{i}_{common}} \quad \textbf{(Definition 2)}$$

By applying definition 2 to equation 2, the tAI ratio of the two genes can be then written as follows:

$$\mathrm{diff}(tAI_A, tAI_B) = \frac{tAI_A}{tAI_B} = \frac{\prod_i tAI^{i}_{common}{}^{f^{i}_{common}} \times \prod_i tAI^{i}_{uniqueA}{}^{f^{i}_{common}} \times \prod_i tAI^{i}_{common}{}^{f^{i}_{uniqueA}} \times \prod_i tAI^{i}_{uniqueA}{}^{f^{i}_{uniqueA}}}{\prod_i tAI^{i}_{common}{}^{f^{i}_{common}} \times \prod_i tAI^{i}_{uniqueB}{}^{f^{i}_{common}} \times \prod_i tAI^{i}_{common}{}^{f^{i}_{uniqueB}} \times \prod_i tAI^{i}_{uniqueB}{}^{f^{i}_{uniqueB}}}$$

(Equation 3)

The first term in equation 3 is equal to one by definition; it represents conservation, and thus was removed from further analyses. Applying the log function on the rest of the terms in equation 3 results in the following decomposition:

$$\log(\mathrm{diff}(tAI_A, tAI_B)) = \log\left(\frac{\prod_i tAI^{i}_{uniqueA}{}^{f^{i}_{common}}}{\prod_i tAI^{i}_{uniqueB}{}^{f^{i}_{common}}}\right) + \log\left(\frac{\prod_i tAI^{i}_{common}{}^{f^{i}_{uniqueA}}}{\prod_i tAI^{i}_{common}{}^{f^{i}_{uniqueB}}}\right) + \log\left(\frac{\prod_i tAI^{i}_{uniqueA}{}^{f^{i}_{uniqueA}}}{\prod_i tAI^{i}_{uniqueB}{}^{f^{i}_{uniqueB}}}\right)$$

(Equation 4)

The first term involves the overlapping codon frequency in both organisms, but different tAIs, thus representing the *trans* component. The second term involves the overlapping tAIs in both organisms, but different codon frequencies, thus representing

the *cis* component, and the third term involves different tAIs and different codon frequencies, representing co-evolution.

2.1.2. tAI decomposition in yeast

I chose to analyze the decomposition of the tAI on eight yeast species, as described in (Man and Pilpel 2007), with the exclusion of *S. bayanus* (for which we do not have a clear estimate of the tRNA pool). I paired each organism with the other seven, creating 28 pairs. For each of the pairs I generated an orthologous genes list using the Inparanoid algorithm (Remm et al. 2001, see Methods), and then calculated the decomposition of the tAI difference for each pair of genes.

To evaluate the behavior of the different components across pairs, I first calculated the distributions of each of the components in each of the 28 pairs. Figure 2 shows the component distributions among 3 pairs: *S. cerevisiae* vs. *K. lactis*, *S. cerevisiae* vs. *Y. lipolytica* and *S. pombe* vs. *Y. lipolytica*. Mean and Standard deviation of all the pairs can be found in appendix 1.

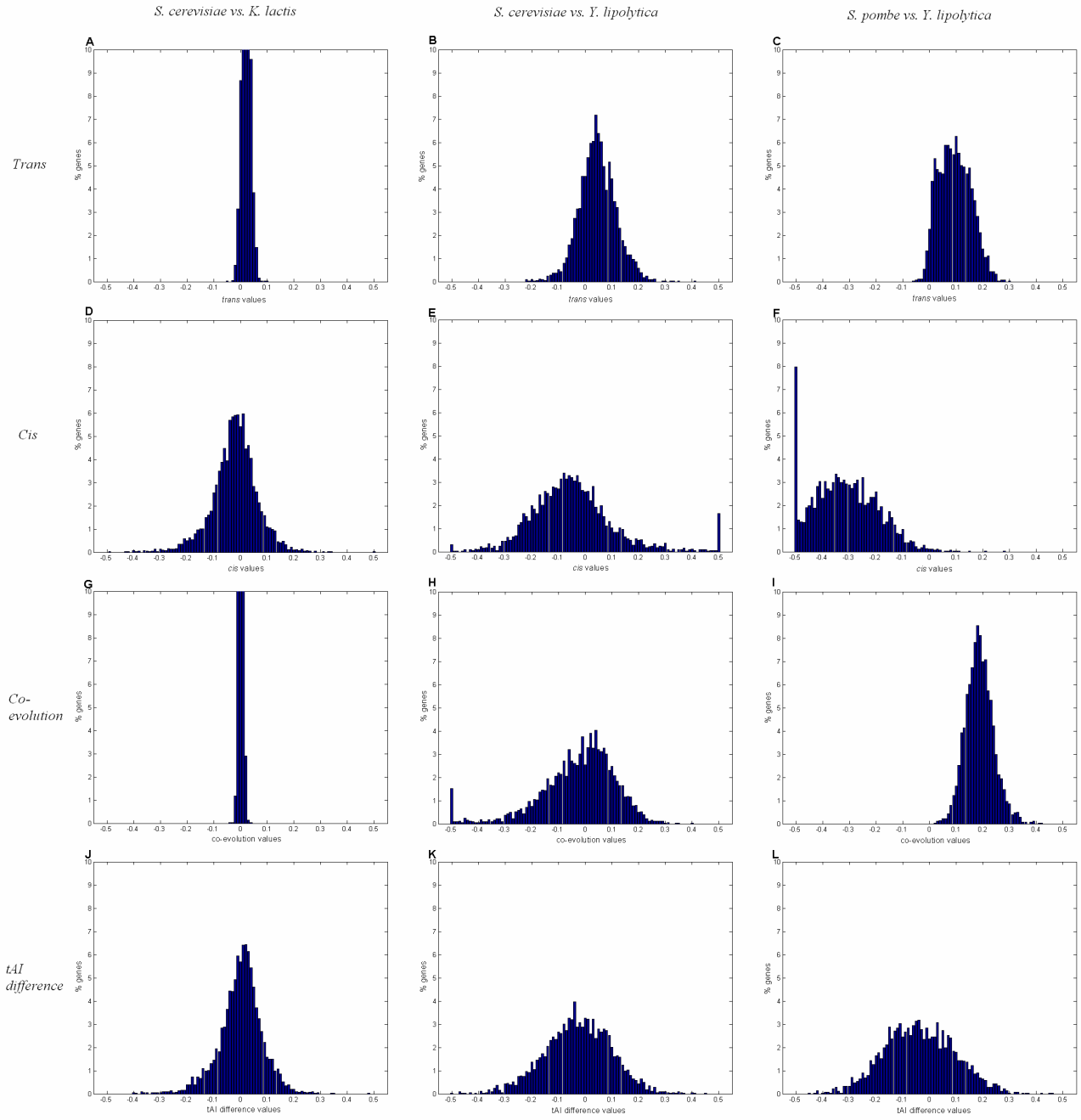


Figure 2: Distribution of the tAI change components. Histograms of the tAI components among 3 pairs are shown. Each column involves one pair of organisms. The pairs are *S. cerevisiae* vs. *K. lactis* in column 1; *S. cerevisiae* vs. *Y. lipolytica* in column 2 and *S. pombe* vs. *Y. lipolytica* in column 3. Each row depicts one component. Row 1, A-C: *trans*. Row 2, D-F: *cis*. Row 3, G-I: co-evolution. Row 4, J-L: the total tAI difference. In an '*X vs. Y*' comparison, negative values imply advantage to *Y* over *X*. Note that the same axes ranges were employed in all the panels; this led in some cases to artifactual jumps in frequencies at the extreme bars (e.g., panel F, the leftmost bar).

One major observation is that the *cis* component plays a significant role in each of the pairs tested. The *cis* components (figure 2, D-F) have a wide distribution of the values among all the pairs tested. On the other hand, the *trans* components (figure 2, A-C)

and co-evolution components (figure 2, G-I) usually have a narrower distribution. In some pairs, *cis* is the only component with a significant contribution to the translation efficiency difference (for example, see figure 2, A, G).

The distributions of the individual components usually show a tendency in favor of one organism over the other. This is clearly shown in extreme cases, in which in almost all genes, one component gives an advantage to a specific organism, while another component gives the advantage to the other one. Such an example is shown in figure 2, in the *S. pombe vs. Y. lipolytica* pair (column 3). While the *trans* and the co-evolution components clearly give advantage in the translation to *S. pombe* genes, the *cis* component gives the advantage to *Y. lipolytica* genes.

Two species that show a consistent and interesting behavior in their components distributions across all comparisons are *Y. lipolytica* and *S. pombe*. The tRNA pools of all the yeasts tested evolved very slowly and are highly correlated. However, the tRNA pool of *Y. lipolytica* shows a significant divergence from most of the other pools (Man and Pilpel 2007). An apparent consequence of this difference is that the *trans* component in every comparison involving *Y. lipolytica* shows a disadvantage to that species; however, the disadvantage appears to be compensated by an advantage in the *cis* component (e.g., figure 2, B, E). *S. pombe* is the furthest organism in the evolutionary tree among all the species in the analysis. Its distance from the hemiascomycotic species [all but *A. nidulans*, see (Souciet et al. 2000)] is 350-1,000 MYA (Berbee and Taylor 2001). When examining the behavior of components in pairs involving *S. pombe* we can see a consistent disadvantage for *S. pombe* in the distributions of the *cis* component, with a compensating effect in the co-evolution component. In some pairs, this behavior is consistent across all genes, meaning that almost, if not all of the genes show this trend (figure 2, F, I). These results suggest that significant changes in one component (e.g. tRNA pool in *Y. lipolytica*) were counter-acted by changes in the second component (e.g. the coding sequences in *Y. lipolytica*) in order to conserve the translation efficiency.

To verify the significance of the results and to alleviate concerns about biases in the decomposition that might generate the observed behavior, I generated a random set of genes and tRNA pools, comprising of 1,000,000 pairs of genes and tRNA pools, and calculated their decomposition. In each pair, the coding sequences and the tRNA genes copy numbers were drawn from a uniform distribution (see Methods).

Figure 3 shows the distribution of each of the components and the tAI difference. As seen in this figure, all components are normally distributed with mean of 0, alleviating the concerns for any bias in the decomposition.

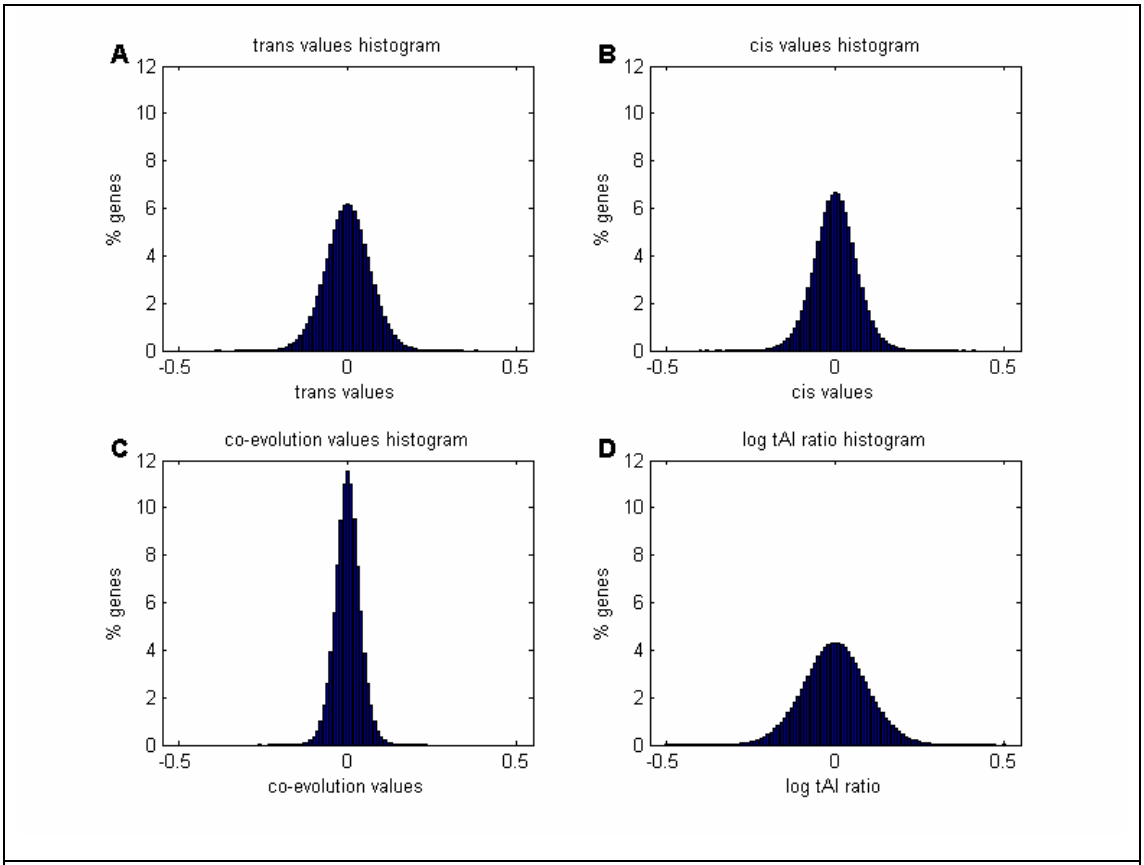


Figure 3: distributions of the tAI and its decomposition components in a randomly generated data. Each histogram contains values from 1,000,000 independent random choices of two genes and two tRNA pools.

In order to create a balanced distribution of the tAI change, an advantage in one component must be compensated with a disadvantage in another component. Since the *trans* component is usually small, this effect would be most noticeable in the relations between the *cis* and co-evolution components. To test this hypothesis, I calculated the correlation coefficient between the *cis* and co-evolution components of the genes in every pair. The results indeed show a significant anti-correlation between the *cis* and co-evolution components, with correlation coefficient ranging from -0.59 to -0.9, and maximal p-value of $2.15*10^{-255}$ (five pairs in which *cis* was the only dominant component were excluded from the calculations). Reassuringly tests of the correlation between the *cis* and co-evolution components in the random data resulted in a much weaker correlation coefficient of -0.47, with $p < 2.23*10^{-308}$. Although this is a weaker correlation, it does suggest a bias in the calculation towards anti-correlation between the *cis* and co-evolution components.

To measure the contribution of each component to the tAI change of the analyzed pairs, and to test the relationships between the components on a global scale, I calculated the pair-wise correlation between the means of the components' distributions. In addition, I calculated the correlations between the means of the components' distributions and the total change in tAI of every pair. Table 1 shows the correlation coefficients and their corresponding p-values. The analysis clearly shows the high correlation between the *cis* components and the tAI changes, and the high anti-correlation between the *cis* components and the co-evolution components.

	Trans	Cis	Co-evolution	tAI change
Trans	-	-0.65	0.41	-0.49
Cis	0.0002	-	-0.95	0.75
Co-evolution	0.03	$2.11*10^{-14}$	-	-0.61
tAI change	0.008	$4.2*10^{-6}$	0.0006	-

Table 1: correlations between the means of the distributions of the tAI change components. The components where calculated for genes in 28 yeast pairs, and a mean value was calculated for each component in every pair, creating, for each component, a vector of 28 means. The correlation was calculated between the vectors. The upper triangle contains the Pearson correlation coefficients, and the lower triangle contains the corresponding p-values.

The above results show that the difference in the tAI change distribution between different pairs can be mostly explained by the *cis* component. This suggests that the differences in the values of the tAI changes between pairs are mostly due to the changes in the coding regions, as opposed to changes due to the tRNA repertoires. The results above show how components act on a global scale. To test what is the components behavior on a per pair basis, I calculated, in each species pair, the Pearson correlation coefficient between each component and the tAI ratio, and between the sum of two components and the tAI ratio, using the entire ortholog set for each pair. I then clustered the resulting correlation matrix using hierarchical clustering to detect species pairs with similar patterns. Figure 4 shows the results of the correlation and clustering analysis using only the 3 components. The clustering clearly divides the pairs into three distinct groups (Figure 4, top panel), which are different in the explanatory power of the components. In the first group (top panel, blue), the component with the highest correlation to the tAI change is the *trans* component (bottom panel, blue). Not surprisingly, this group contains only pairs which involve one hemiascomycotic species and the organism whose tRNA pool is the most divergent from them, namely *Y. lipolytica*. The second group (top panel, red) includes pairs in which the *cis* component (bottom panel, green) is highly correlated to the tAI change, while the *trans* component has a relatively low correlation. Three

of the species in this group (*S. cerevisiae*, *K. lactis* and *C. glabrata*) have a very high correlation between their tRNA pools, suggesting an explanation for the small contribution of the *trans* component to the tAI change. However, *D. hansenii* and *C. albicans* do not have a much similar tRNA pool, suggesting that the minor contribution of the *trans* component is due to other reasons. The third group, which includes the majority of the pairs, contains pairs with relatively high correlation of both the *cis* and *trans* components.

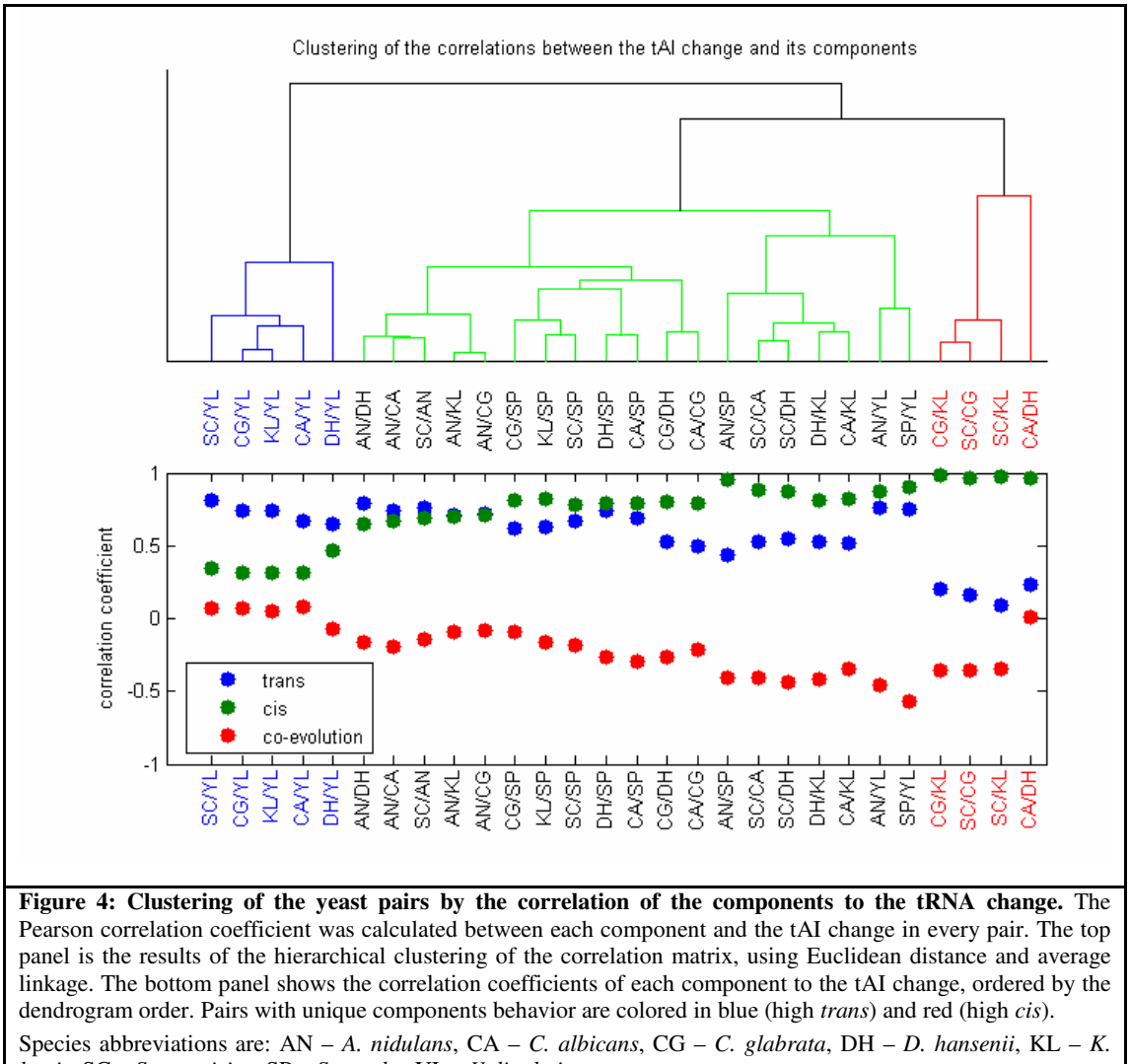


Figure 4: Clustering of the yeast pairs by the correlation of the components to the tRNA change. The Pearson correlation coefficient was calculated between each component and the tAI change in every pair. The top panel is the results of the hierarchical clustering of the correlation matrix, using Euclidean distance and average linkage. The bottom panel shows the correlation coefficients of each component to the tAI change, ordered by the dendrogram order. Pairs with unique components behavior are colored in blue (high *trans*) and red (high *cis*).

Species abbreviations are: AN – *A. nidulans*, CA – *C. albicans*, CG – *C. glabrata*, DH – *D. hansenii*, KL – *K. lactis*, SC – *S. cerevisiae*, SP – *S. pombe*, YL – *Y. lipolytica*.

Figure 5 shows the clustering of the pairs when taking into account the correlations between pairs of components to the tAI change. Adding an additional component to the *cis* component improves the correlation to the tAI change, with *cis*+co-evolution (figure 5, bottom panel, yellow) being usually the highest correlated component and the *trans*+*cis* components (cyan) are almost as high. Clustering of the complete matrix gave a slightly different order of the pairs than clustering by 3 components. While the

pairs involving *Y. lipolytica* and one hemiascomycotic species still retain their uniqueness by being clustered together, the other clusters reflect the evolutionary relationships between the species, with pairs involving a hemiascomycotic species with either *A. nidulans* or *S. pombe* (both are not hemyascomycotic) being grouped in one cluster (figure 5 top panel, red), and pairs involving species from the same evolutionary branch (excluding *Y. lipolytica*) are clustered together (top panel, green). This cluster is characterized by low correlation of the *trans*+co-evolution components.

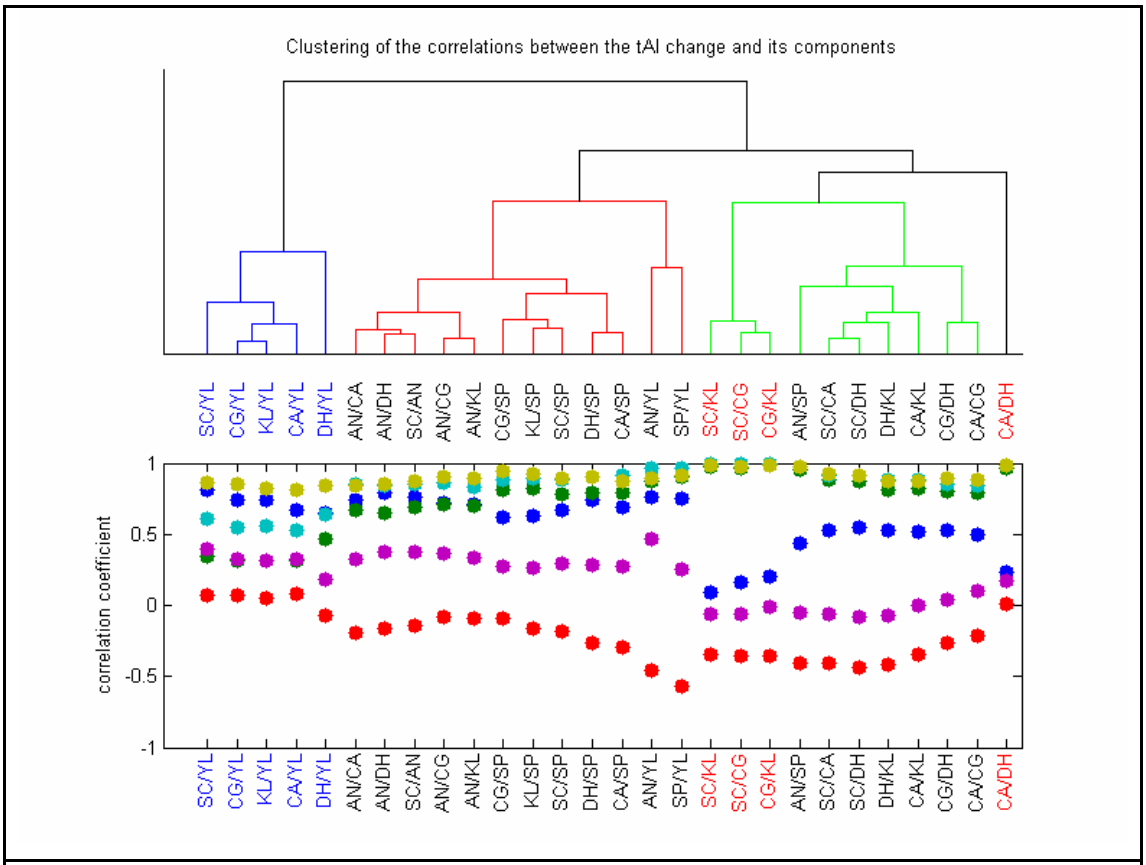


Figure 5: Clustering of the yeast pairs by the correlation of the components to the tRNA change. The procedure was the same as in figure 3, with the addition of the components pairs: yellow - *cis*+co-evolution, cyan – *trans*+*cis* and purple – *trans*+co-evolution.

2.1.3. Functional analysis of the tAI decomposition

Given the tAI decomposition of the 28 yeast pairs, I could turn to examine the relationships between gene functions and decomposition patterns by cluster analysis. Clustering is commonly used in gene expression analysis to identify co-expressed genes and generate hypotheses about their involvement in the conditions tested (Boutros and Okey 2005). Analogously, I used k-means clustering to find sets of similarly decomposed genes (see Methods). In this analysis genes that show similar pattern of tAI decomposition are grouped together. Figure 6 show the result of assigning the decomposition of the *S. cerevisiae vs. Y. lipolytica* pair into 30 clusters.

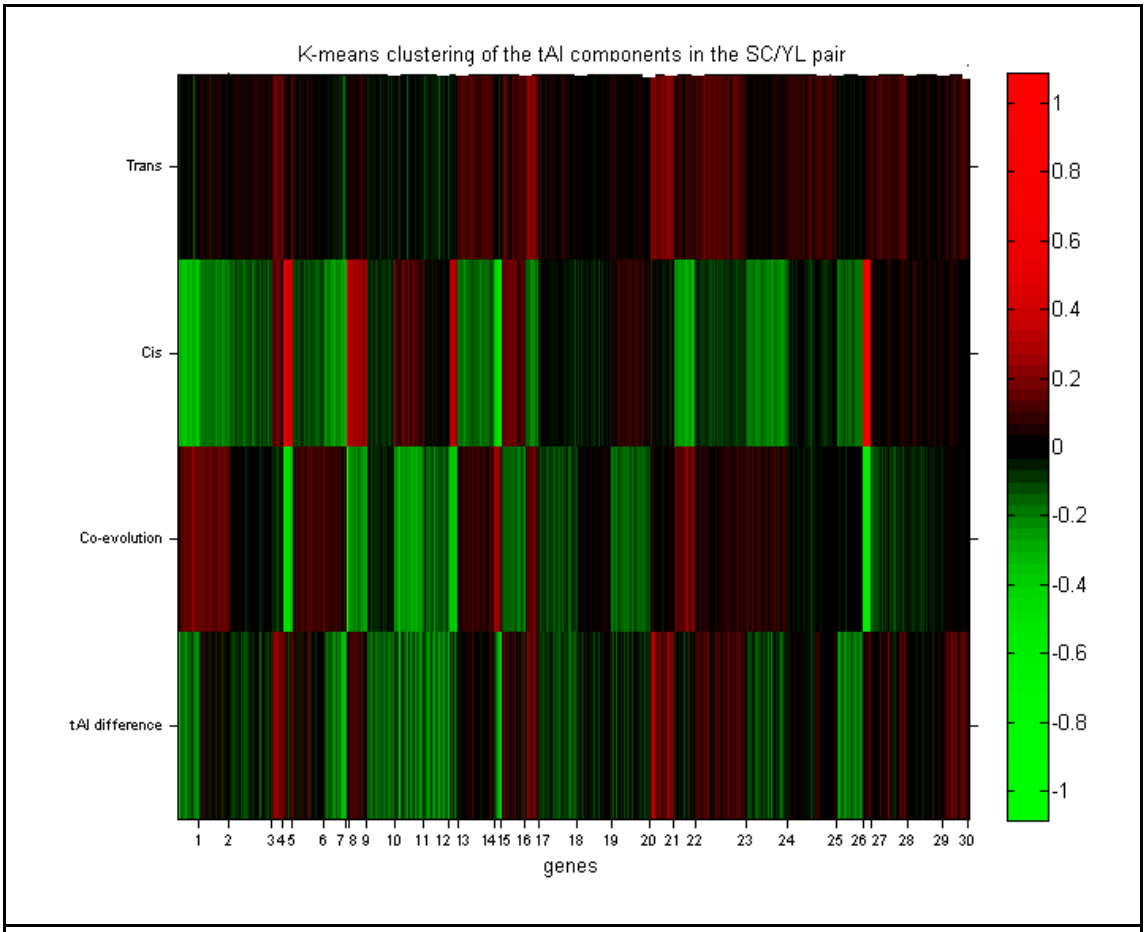


Figure 6: k-means clustering of the tAI decomposition for *S. cerevisiae* and *Y. lipolytica*. Genes in the k-means clustering were assigned into 30 clusters using Euclidean distance (cluster numbers are indicated below, and to the right of each cluster). Starting points were chosen randomly 500 times and the optimal clustering result is displayed. The color code indicates the direction of the advantage in terms of translation efficiency. Red means an advantage to ***S. cerevisiae*** over ***Y. lipolytica*****; green implies the opposite.**

As shown in figure 6, we can find clusters that represent many decomposition patterns. Some clusters represent dominancy of only one component, either *trans* (figure 6, cluster 21), *cis* (cluster 26) or co-evolution (cluster12), and some represent an interplay between more than one component, (cluster 10), or more complex cases where the various components counter-act each other minimizing the net tAI difference (cluster 2).

In order to shed light on phenotypic differences that might be implied by the tAI decomposition, I looked for enrichments of functional terms from the Gene Ontology (GO) database (Harris et al. 2004) in each cluster, for every pair of species. I was specifically interested in genes related to the glucose repression phenotype. Glucose repression is the preference of metabolizing glucose through fermentation rather than respiration even under aerobic conditions (Barnett and Entian 2005). Man and Pilpel (2007) found a significant difference in the translation efficiency in genes related to this phenotype between yeast species that display the glucose repression phenotype

and those which do not. Specifically, they found that glycolytic genes are translated more efficiently in yeast species that show this phenotype, including *S. cerevisiae, C. glabrata* and *S. pombe*, while aerobic respiration genes are translated more efficiently in yeast species that do not show this phenotype. The same behavior was detected in the relations between the cytosolic ribosomal proteins (CRPs) which are translated more efficiently in species that show the phenotype, and the mitochondrial ribosomal proteins (MRPs), which are translated more efficiently in species that do not show this phenotype. Yet whether these differences in tAI arise due to change in *cis*, *trans*, or co-evolution of the tRNA pool and coding sequences, was so far not known. To analyze the decomposition schemes of those genes I extracted the clusters which are enriched for the related categories from all pairs where one yeast species shows the phenotype and the other does not. Figure 7 shows the clusters which are enriched for the 4 related GO categories in selected pairs.

Examination of these results shows that there is heterogeneity across the species pairs with respect to the relative contribution of each component. For example, in the clusters enriched for the "organellar ribosome" category (which mainly corresponds to mitochondrial ribosomal proteins, see Methods), we can observe clusters with dominancy of the *cis* component (figure 7, second panel, *S. cerevisiae vs. K. lactis*), or co-evolution component (second panel, *S. cerevisiae vs. Y. lipolytica*). *Trans* dominant clusters appear only in the "glycolysis" category clusters (third panel, *S. cerevisiae vs. C. albicans* and *S. cerevisiae vs. K. lactis*) or in the "cytosolic part" category (which mainly corresponds to cytosolic ribosomal proteins) clusters (first panel, *S. cerevisiae vs. C. albicans*, *S. cerevisiae vs. K. lactis* and *C. glabrata vs. K. lactis*). The existence of a *trans* dominant cluster in the *S. cerevisiae vs. K. lactis* pair is very interesting, since those organisms have a highly correlated tRNA pool, and the tAI change is dominated by the *cis* component, when examining all the genes (see figure 2, column 1, and figure 4). This indicates that while the *cis* component is mainly dominant in explaining tAI difference between *S. cerevisiae* and *K. lactis*, the *trans* component is dominant among CRP genes and glycolysis genes which owe their enhanced tAI in *S. cerevisiae* to a trans effect.

Although we do not see a consistent behavior in each category, there is, to some extent, a consistent behavior across categories. This can be viewed in the example above, where both the "cytosolic part" category and "glycolysis" category genes are

enriched in *trans* dominant clusters in *S. cerevisiae vs. C. albicans* and *S. cerevisiae vs. K. lactis* pairs. On the other hand, the *S. pombe vs. Y. lipolytica* pair shows an inconsistent behavior, where the "glycolysis" category genes (third panel, last group) show a high *trans* component giving the advantage to *S. pombe* genes, while the "cytosolic part" category genes do not show this feature, giving the advantage to the *Y. lipolytica* genes.

Figure 7: tAI decomposition of clusters enriched for the glucose repression related categories. Each plot displays clusters enriched for a specific category related to the glucose repression phenotype, taken from pairs where one organism show the phenotype ad the other does not. Number of genes and enrichment strength can be found in appendix 2. Species abbreviations are: AN – *A. nidulans*, CA – *C. albicans*, CG – *C. glabrata*, DH – *D. hansenii*, KL – *K. lactis*, SC – *S. cerevisiae*, SP – *S. pombe*, YL – *Y. lipolytica*. Note that each pair can contain more than one cluster. The color code indicates the direction of the advantage in terms of translation efficiency. Red means an advantage to a glucose repression capable organism (SC, CG or SP) over the non capable organism; green implies the opposite.

Another category that was of interest to me is the DNA repair genes. Man and Pilpel found that genes in *Y. lipolytica* and *D. hansenii* that are involved in DNA repair processes have a significantly lower translation efficiency compared to their orthologs in the other yeast species tested. Reassuringly, a preliminary experimental work performed in our lab suggested that *D. hansenii* might have a reduced tolerance to UV radiation compared to *S. cerevisiae*, consistent with lower translation efficiency of the DNA damage response in *D. hansenii* (Yuval Dorfan, unpublished data). Figure 8 shows the clusters enriched for DNA repair genes. An interesting observation is the

strong part played by the *trans* component in all the enriched clusters to the extent of being the main force that underlie the total change in the *S. cerevisiae vs. D. hansenii* pair. Specifically, the analysis suggests that the enhanced translation efficiency of the DNA repair enzyme genes in *S. cerevisiae* is due to changes in the tRNA pool.

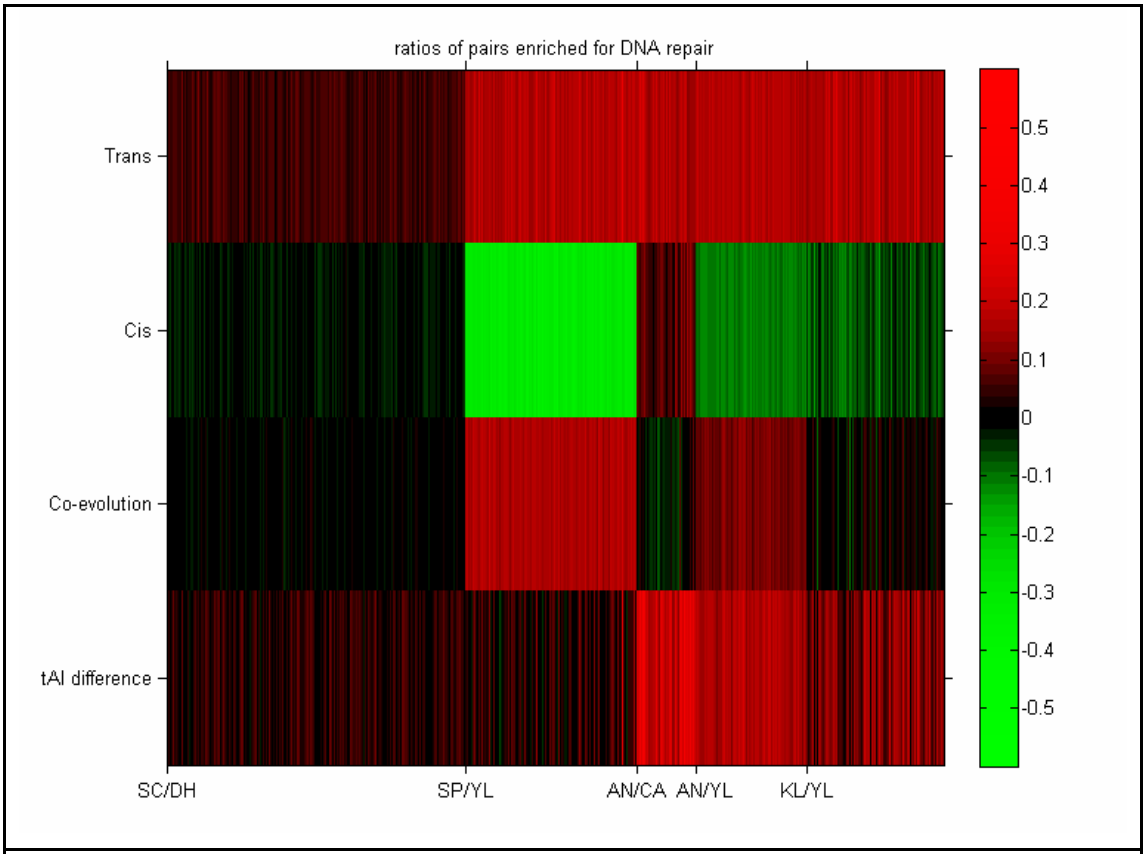


Figure 8: tAI decomposition of clusters enriched DNA repair. The plot displays clusters enriched for DNA repair related genes. The number of genes and enrichment strength are as follows: SC/DH (22 genes out of 212 in the cluster, 119 in the category, $p < 7.7*10^{-7}$) SP/YL (18 genes out of 122 in the cluster, 90 in the category, $p < 1.04*10^{-7}$) AN/CA (8 genes out of 42 in the cluster, 96 in the category, $p < 5.04*10^{-5}$) AN/YL (14 genes out of 79 in the cluster, 91 in the category, $p < 1.6*10^{-7}$) KL/YL (13 genes out of 98 in the cluster, 105 in the category, $p < 1.56*10^{-5}$). The color code indicates the direction of the advantage in terms of translation efficiency. Red means an advantage to the left side organism in a pair; green implies the opposite.

Trans dominant clusters as appear in the analyses in figures 7 and 8, pose as very interesting candidates for further analysis, since they may suggest that a shift in the tRNA pool directly influenced the lifestyle of the organisms that undergone this shift.

2.2. The effect of viral tRNA genes on the translation efficiency of viruses and their hosts

In this part, which was done in collaboration with Keren Limor Waisberg and Prof. Avigdor Scherz, I chose to analyze the effect of viral tRNA genes using *cyanobacteriophage* Syn9. Syn9 is a large phage related to T4 which infects cyanobacteria. Its genome contains 226 protein coding genes and six tRNA genes (Peter R. Weigele 2007), which make this virus an interesting test case for the present purposes. In addition, Syn9 was found to be able to infect a wide variety of hosts (Sullivan et al. 2003) from two different genera, the *Synechococcus* and *Prochlorococcus* (Rocap et al. 2002), making it a good candidate for comparative analysis.

In order to analyze the effect of the Syn9 virus tRNA genes on the translation efficiency of its own genes and of the genes of various potential hosts, I chose 12 different cyanobacteria which have completely-sequenced genomes. Syn9 was previously isolated from *Synechococcus sp.* WH 8012 (Waterbury and Valois 1993), whose genome is not yet completely sequenced. Its closest kin with a completely sequenced genome is *Synechococcus sp.* WH 8102 which is one of the 12 bacteria analyzed in this study. In addition to it, one more bacterium from the *Synechococcus* genera was analyzed (WH7803) and another 10 from the *Prochlorococcus* genera, of which six are *LL-Prochlorococcus* (MIT9313, MIT9303, SS120, NATL1A, NATL2A and MIT9211) and 4 are *HL-Prochlorococcus* (MIT9515, MED4, MIT9215 and MIT9312). In cross infection experiments, Syn9 was found to successfully infect all but NATL1A, SS120, MIT9215 and MIT9312 (Sullivan et al. 2003).

2.2.1. Differences in the tRNA repertoire among the cyanobacteria

Using a HMM-based approach (Lowe and Eddy 1997) I identified all tRNA coding genes, in the 12 bacteria and the cyanophage (Appendix 3). The number of tRNA genes in the selected bacteria ranges from 37 genes in the *HL-prochlorococcus* bacteria to 43 genes in the *Synechococcus* bacteria. 2 *LL-Prochlorococcus* bacteria, MIT9313 and MIT9303 have similar tRNA genes as the *Synechococcus* bacteria, and they are closest to them evolutionarily (Rocap et al. 2002). tRNA genes which are absent in the *HL-prochlorococcus* but present in the *Synechococcus* and the 2 mentioned *LL-Prochlorococcus* bacteria have anti-codons for the codons CTC and

CTG (Leucine), CCG (Proline), GTG (Valine), GCG (Alanine) and GGG (Glycine). Two tRNA genes have additional copies, those include genes that correspond to codons ATC (Isoleucine) and GCA (Glycine), one gene, which corresponds to the ATG codon (Methionine), has additional copy in the *HL-prochlorococcus* bacteria and one tRNA gene that corresponds to the CTT codon (Leucine) does not appear in the *Synechococcus* bacteria but appears in the *HL-prochlorococcus* bacteria. For a complete table of tRNA gene copy numbers see appendix 3.

All of the tRNA genes which are unique to the *Synechococcus* bacteria correspond to codons with high GC content. This is no surprise, taking into consideration the high GC content of the bacterial genomes. While the *HL-prochlorococcus* bacteria have an average GC content of 30%, most of the *LL-Prochlorococcus* bacteria have an average GC content of 37%, the MIT9313 and MIT9303 have an average GC content of 50% and the *Synechococcus* bacteria have an average GC content of 60%. The six tRNA genes encoded by the virus correspond to the codons: TTA (Leucine), ACA (Threonine), AAC (Asparagine), AGA (Arginine), GTA (Valine) and GCA (Alanine). The tRNA gene for GCA is the only gene encoded by the virus with more than one copy in the host genome. While all the other 5 tRNA genes exist in one copy on each host genome, the GCA tRNA gene has two copies in the two *Synechococcus* bacteria and the closely related two *LL-prochlorococcus*. Most of the tRNA genes encoded by the virus correspond to for low GC codons, in accordance with the GC content of the virus genome which is about 40%.

2.2.2. The effect of SYN9 tRNA genes on translation efficiency

Adding tRNA genes to the cellular pool can either improve or decrease the tAI value of genes (see Discussion for details). In order to test the effect of adding the virus tRNA genes to the pool of the host tRNA genes on both the virus genes and the corresponding host genes, I calculated, for each pair of virus-host, what would be the tAI value for each gene if the virus would not carry any tRNA gene, and what is the actual tAI value of every gene, with the combined pool of the viral and bacterial tRNA gene pool. For each gene I then calculated the log ratio of the tAI values, and tested whether the addition of the tRNA genes improved or decreased the gene's tAI value.

2.2.2.1. The effect of SYN9 tRNA genes on its own translation efficiency

Figure 9 shows the effect of the tRNA genes on every viral gene in the 12 bacteria. It is clearly seen, that most of the genes are up regulated in the background of *Synechococcus* bacteria and to a less extent in the high-GC content *LL-prochlorococcus*, while most of the genes are down regulated in the background of the low-GC content *LL-prochlorococcus* and *HL-prochlorococcus*. This suggests that the viral tRNA genes are adapted to work in the *Synechococcus* genomic background.

To check whether the characteristics of the changes are different between bacteria which are susceptible to the virus infection and those who are not, I clustered the tAI ratio matrix according to the change profile of each host. The clustering process divided the bacteria into two groups: One containing the four bacteria in which the viral tRNA genes have mostly a positive effect on the tAI, and another, containing the eight bacteria in which the viral tRNA genes have mostly a negative effect. The latter group contains all the resistant bacteria. However, the tAI changes are not sufficient to explain the infection pattern, since the resistant bacteria did not cluster together, but rather with their evolutionary closets kin. Note, however, that while the MIT9303 and MIT9313 are evolutionary related to the eight bacteria that show a negative effect, the clustering grouped them with the *Synechococcus* bacteria, due to the positive effect of the viral tRNA genes.

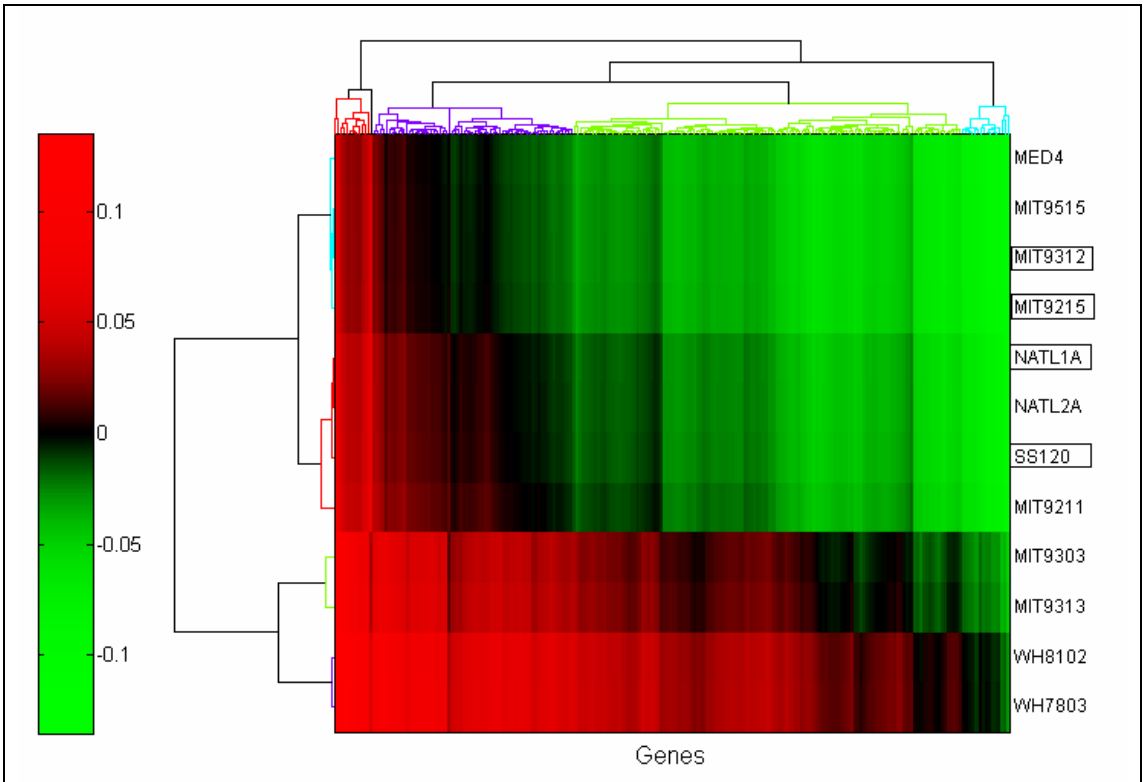


Figure 9: The difference in tAI of the viral genes in the presence, or absence, of the viral tRNA.
The colors represents the log ratio of the tAI when using the viral tRNA and the tAI with only the host tRNA. Red values correspond to improvement in the tAI; green values correspond to decrease in the tAI.
Each column represents a gene, and each row represents a different bacterium. The values were clustered according to the organism profiles (rows) and then according to the genes profiles (column)
A box around an organism name indicates resistance to infection by the virus

I have focused further on the WH8102 genome – the host in which the viral genes showed the highest extent of improvement in tAI due to the addition of the viral tRNAs.

The change to the tAI of the viral genome in WH8102 as a host ranges from 3% decrease to 14% increase; the average change is 4% increase. I wanted to test if viral genes which belong to different functional groups respond differently to the presence of the viral tRNA genes. This would suggest that the viral tRNA genes were selected to optimize specific groups. For this test, I calculated the change in tAI for viral genes belonging to different functional groups classified according to (Peter R. Weigele 2007). The virus proteins were classified into DNA metabolism proteins (22 proteins), Transcription and translation regulation (3 proteins), Photosynthesis proteins (7 proteins), Virion proteins (29 proteins) and other known function proteins (12 proteins). The rest of the proteins (153 proteins) were unclassified. The behavior of the functional group was tested with WH8102 as the host bacteria. The box plot of the ratios of the tAIs with and without the virus tRNA genes shows that the group with

the highest change is the virion protein group (figure 10). Virion proteins are the proteins that assemble the virus envelope. Each protein appears in multiple copies in every virus particle - sometimes even reaching 1000 copies per one virus (Mesyanzhinov et al. 2004) – suggesting a high translation efficiency for those genes, in accordance to the observation in figure 10.

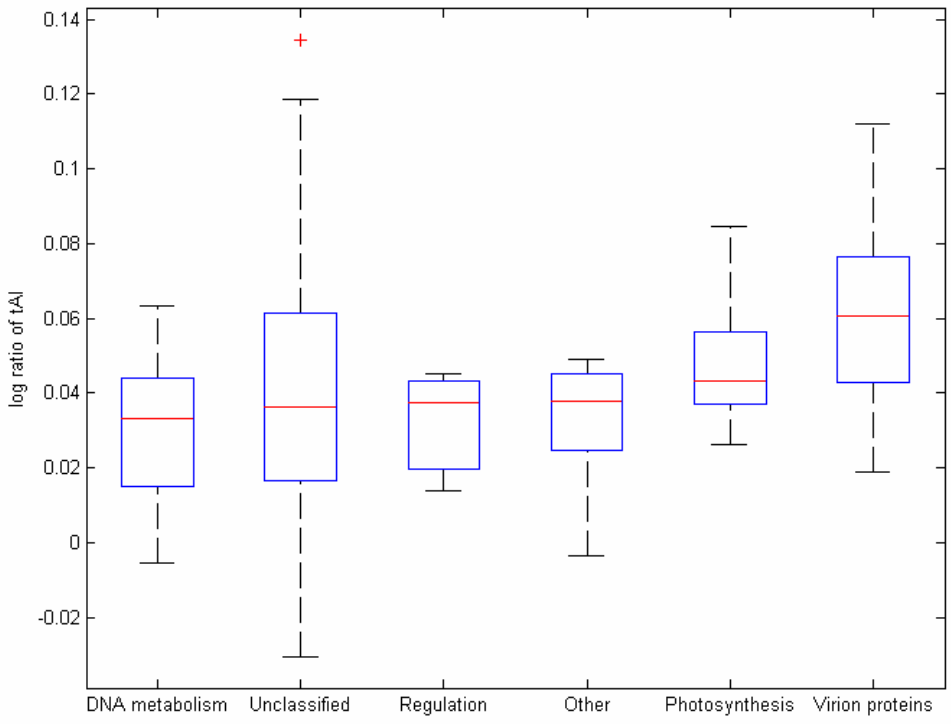


Figure 10: The tAI change of different functional groups in the viral genome. A box plot of the log ratio of tAI is shown. Groups are ordered in an ascending order of the median (red line).

One possible explanation for the wide span of enhancement levels among the various virus genes is their initial tAI - we speculated that perhaps genes with low tAI without using the virus tRNA genes are the ones which gain the most improvement. In order to test this hypothesis I calculated the Pearson correlation between a gene's tAI value when only the host tRNA pool is available, and the improvement to the tAI value with the viral tRNAome available. The correlation coefficient found is weak and implying the opposite direction (0.17, p-value < 0.012), namely that genes with high tAI without using the virus tRNA gain the most improvement.

2.2.2.2. The effect of SYN9 tRNA genes on the host ‘s translation efficiency

To test the different behavior of the hosts' genomes when adding the viral tRNA genes, I calculated the distribution of the tAI ratio of all ORFs in all 12 bacteria. While the tAI of *Synechococcus* WH8102 genes is decreased on average, as do the

tAI of WH7803 and the two high-GC content *LL-prochlorococcus*, the tAI of the other bacteria remain unchanged on average (figure 11). I used the Kruskal-Wallis test (Rice 1995), to check for differences in the distributions of the tAI ratios, and found them to be significantly different. Utilizing *post-hoc* Wilcoxon rank-sum tests, after correcting for multiple hypotheses testing (see Methods), I was able to divide the hosts into 3 different populations which correspond to their taxonomy. The groups are *Synechococcus*, *LL-prochlorococcus* and *HL-prochlorococcus*, with significant differences between the groups, and insignificant differences within each group.

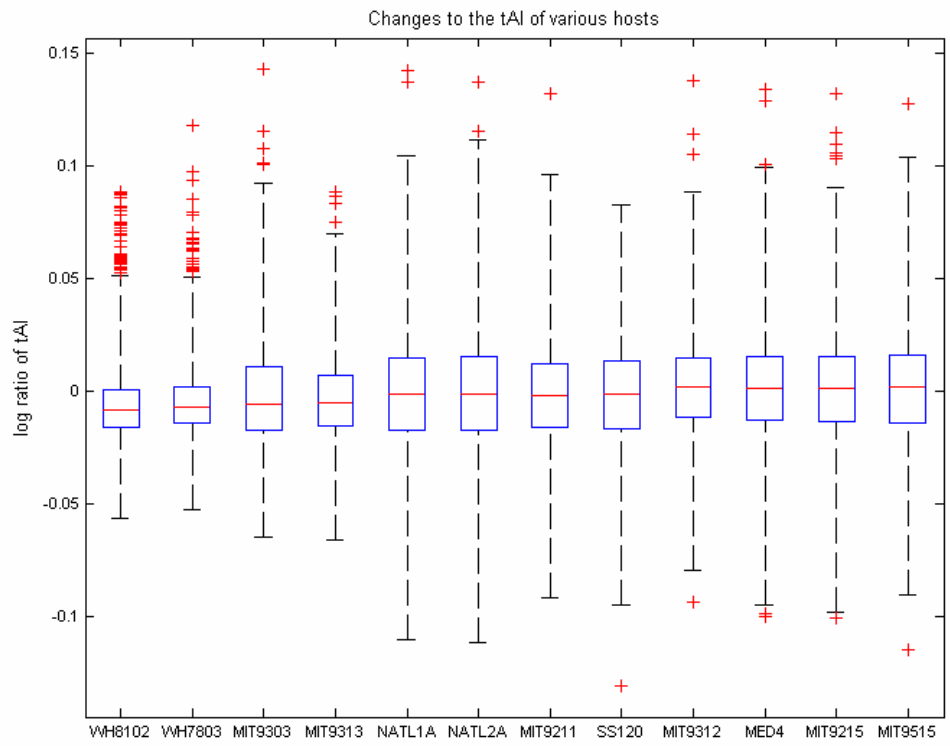


Figure 11: Changes to the tAI of various hosts. A box plot of the log tAI ratio is shown. Organisms are ordered by their evolutionary relations.

It is interesting to see that although the average tAI value of a gene tend to decrease, there is a significant number of outlier genes, whose tAI is improved when introducing the virus tRNA genes to the genome (figure 11, denoted by +). First, I wanted to check the connection between a gene's function and the change to its tAI value. For that purpose, I classified the WH8102 genome into 20 functional groups using data from the CMR website (Peterson et al. 2001). Figure 12 shows the distribution of the changes in the tAI value across the different groups.

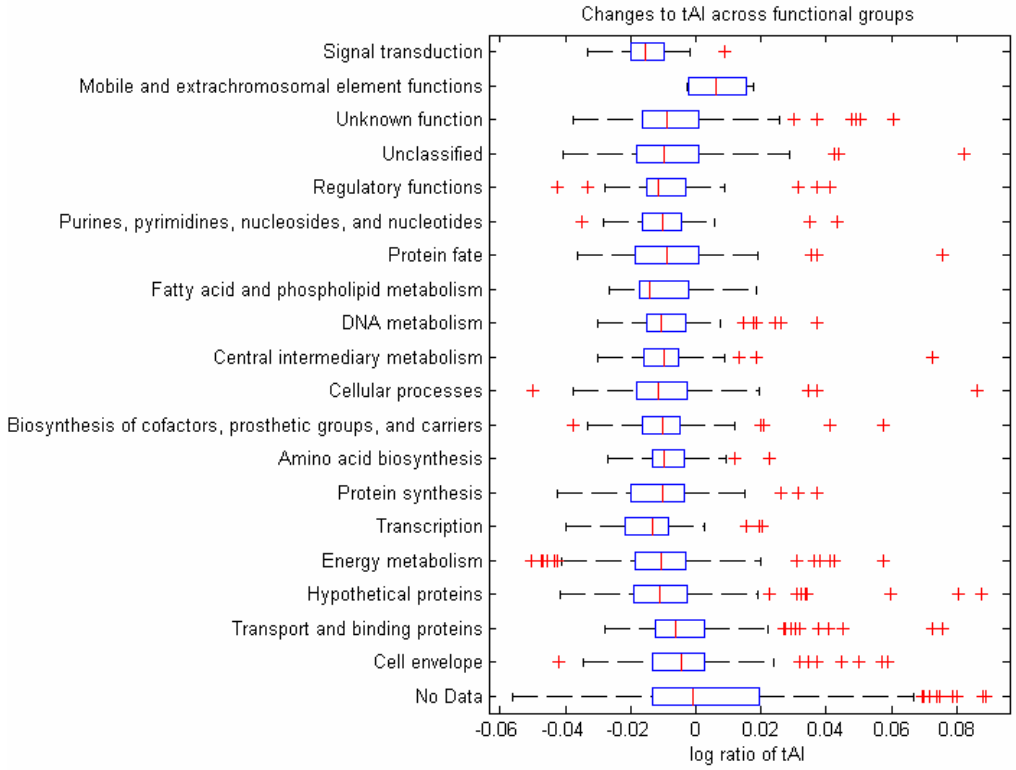


Figure 12: Changes to the tAI across functional groups of the WH8102 genome. A box plot of the log tAI ratio is shown. "No Data" proteins are hypothetical proteins with no significant homology to proteins in other organisms. "Hypothetical Proteins" are hypothetical proteins with homology to hypothetical proteins in other organisms. "Unknown function" proteins are proteins with significant homology to proteins in other organisms, with unknown function. "Unclassified" proteins are proteins not assigned a classification in the CMR database (Peterson et al. 2001).

Statistical analysis of the functional groups shows some interesting results (The groups "Signal transduction" and "Mobile and extrachromosomal element functions" were excluded from the analysis due to the very small number of proteins in these groups). Figure 13 shows the results of the analysis. Proteins classified as "No Data" have mean ranks significantly higher than all of the other functional groups but the "Cell envelope" and the "Transport and binding proteins" groups. The "No Data" proteins mainly include hypothetical proteins with no similarity to proteins in other organisms and with no similarity to defined protein motifs. The "Protein synthesis", "Transcription", "Energy metabolism", and "Hypothetical proteins" groups have mean ranks significantly lower than the "No Data", "Cell envelope" and "Transport and binding proteins" groups. All other comparisons did not yield significant results.

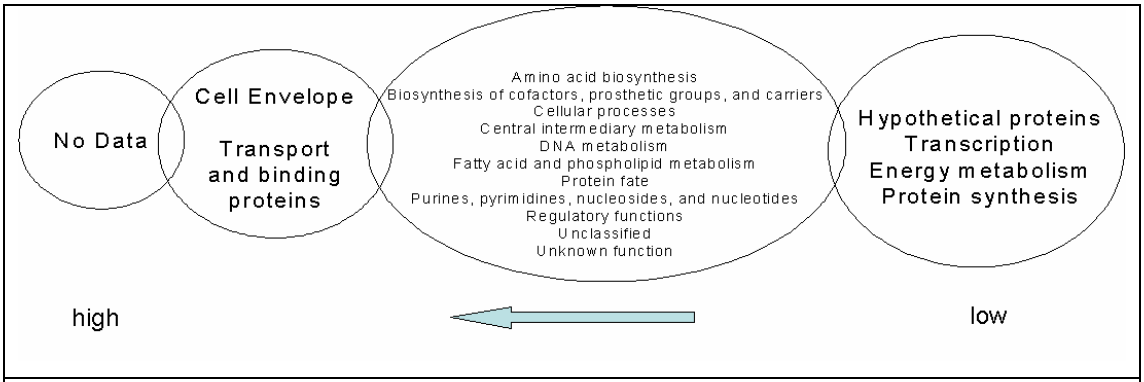


Figure 13: multiple comparisons of mean ranks between functional groups. Comparisons between categories in the same circle or in neighboring circles did not yield significant results.

In order to further test and characterized the genes that show an elevation in their tAI level upon introducing the viral tRNAome, I extracted a list of the most elevated genes (top 200 genes, log ratio ranges from 0.022 to 0.089). Table 2 shows the assignment of the entire genome and the top 200 genes into functional groups. While in the entire genome, 23% of the proteins belong to the "No Data" category, in the top 200 genes 70% of the proteins belong to this category. The large amount of functionally uncharacterized genes might mask the changes in the distribution of genes among the genes with known function. In order to check the distribution only among the genes with assigned functions, I removed from the analysis the four categories which imply unknown function (“No Data”, “Unclassified”, “Hypothetical proteins” and “Unknown function”). Thus I remained with 1274 genes with assigned function (out of 2519) of which 37 are highly improved. The distribution of these genes can be found in table 2. To test the significance of the difference in the distribution, I conducted an enrichment analysis for all the categories present in the highly improved population and found significant enrichment of the "Cell envelope" category (p value < 0.025) and "Transport and binding proteins" category (p value < 0.005). Tests were conducted using the hyper geometric distribution and corrected for multiple hypothesis testing with FDR of 20%).

Category	Entire genome	Top 200 genes	% in Entire Genome (excluding uncharacterized genes)	% in Top 200 genes (excluding uncharacterized genes)
No Data	588 (23.3%)	139 (69.5%)	-	-
Cell envelope	151 (6%)	9 (4.5%)	11.9	24.3
Transport and binding proteins	140 (5.6%)	10 (5%)	11.0	27.0
Amino acid biosynthesis	77 (3%)	1 (0.5%)	6.0	2.7
Biosynthesis of cofactors prosthetic groups, and carriers	123 (4.9%)	2 (1%)	9.7	5.4
Cellular processes	71 (2.9%)	3 (1.5%)	5.6	8.1
Central intermediary metabolism	67 (2.7%)	1 (0.5%)	5.3	2.7
DNA metabolism	82 (3.3%)	3 (1.5%)	6.4	8.1
Fatty acid and phospholipid metabolism	31 (1.2%)	0	2.4	0.0
Mobile and extrachromosomal element functions	5 (0.2%)	0	0.4	0.0
Protein fate	100 (4%)	3 (1.5%)	7.8	8.1
Purines, pyrimidines, nucleosides, and nucleotides	51 (2%)	2 (1%)	4.0	5.4
Regulatory functions	60 (2.4%)	3 (1.5%)	4.7	8.1
Signal transduction	10 (0.4%)	0	0.8	0.0
Unclassified	106 (4.2%)	8 (4%)	-	-
Unknown function	184 (7.3%)	9 (4.5%)	-	-
Hypothetical proteins	377 (15%)	8 (4%)	-	-
Transcription	37 (1.5%)	0	2.9	0.0
Energy metabolism	253 (10%)	6 (3%)	19.9	16.2
Protein synthesis	136 (5.4%)	3 (1.5%)	10.7	8.1

Table 2: Distribution of genes into functional categories in the entire genome and the top 200 elevated genes in WH8102. Assignment of genes into functional categories is based on (Peterson et al. 2001). Percentage may sun up to more than 100% due to some overlaps between the categories.

When looking at the distribution of functional categories of the top 200 elevated genes, we see that more than 70% of them are uncharacterized. Genes that are particularly enhanced by the viral tRNAs might be genes that are relatively new to the host genome, and that may have a viral origin. Such genes are often annotated with "uncharacterized functions" (Daubin and Ochman 2004). It is thus possible that many of the top influenced host genes are of viral origin and hence is the high tAI fold ratio, and the un-characterized function. These genes might not have a role in the host life cycle nor even be transcribed. To try and confront this hypothesis, I turned to check the GC content of the elevated genes. Genes from viral origins tend to be AT rich (Rocha and Danchin 2002). In contrast, the WH8102 genes tend to be GC rich, with an average GC content of 60%. Thus a simple GC content analysis of genes was shown to be a reliable method to detect the origin of host genes (Lawrence and Ochman 1998). Figure 14 shows the GC content of the Syn9 genome, the WH8102 genome, and the top 200 elevated genes. While the viral genome GC content peaks at 40%, and the bacterial genome peaks at 60%, the elevated genes display a bi-modal distribution, with peaks around 40% and 55%. This suggests that the host genes that

are elevated by the viral tRNAs might represent two populations – genes of viral origin, and endogenous bacterial genes. The high GC content group contains 79 proteins of which 62 are uncharacterized. Each of the "Amino acid biosynthesis", "Cell envelope"," Protein fate" and the "Purines, pyrimidines, nucleosides, and nucleotides" groups are represented by one gene; the groups "DNA metabolism", "Protein synthesis" and "Regulatory functions" are represented by two genes each; the "Transport and binding proteins" group is represented by three genes; finally the "Energy metabolism" group is represented by four genes. In addition, out of the 200 genes, only 58 genes reside in low GC content regions associated with phage integrases (Palenik et al. 2003), which are marked characteristics of gene islands that were acquired from phages (Groisman and Ochman 1996). 52 other genes reside in low GC regions without association to phage integrases, while 90 genes do not reside in any of those regions, further substantiating the hypothesis that not all elevated genes come from viral origins.

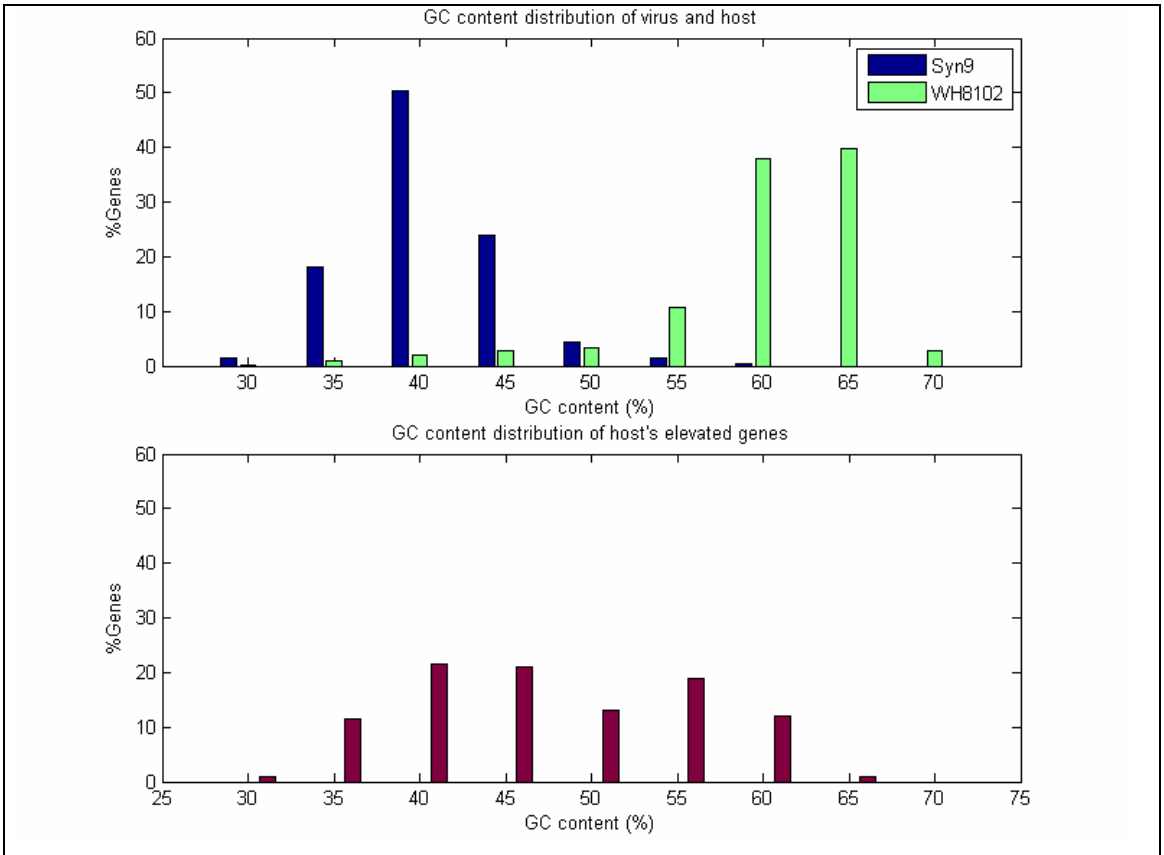


Figure 14: GC content distribution of the virus genes, host genes and elevated genes. Histogram of the GC content (in percent) of Syn9 coding sequences (blue), WH8102 coding sequences (green) and the 200 genes with the most improved tAI (red) are shown.

2.2.3. Optimality of the chosen tRNA genes

The six viral tRNA genes appear to have a significant effect on the translation efficiency of both virus and host genes. Theoretically, there are 54 tRNA genes available for the virus to choose from, and this leads to approximately 25 billion combinations of six tRNA genes. It is interesting to see how optimal is the chosen tRNA set of the virus in terms of changes to the translation efficiency of both viral and host genes. At first, I wanted to see if other sets of six tRNA genes could create the observed extent of elevation to the viral genes' tAI. To test this, I generated 10,000 random sets of six tRNA genes, and measured the mean change each set incurred on the viral tAI. Figure 15 shows the histogram of the measured mean; the red arrow depicts the mean incurred by the actual gene set. It is clearly seen that the chosen set is non-random, as it is one of the best available sets, with only 3% of the random sets with higher mean.

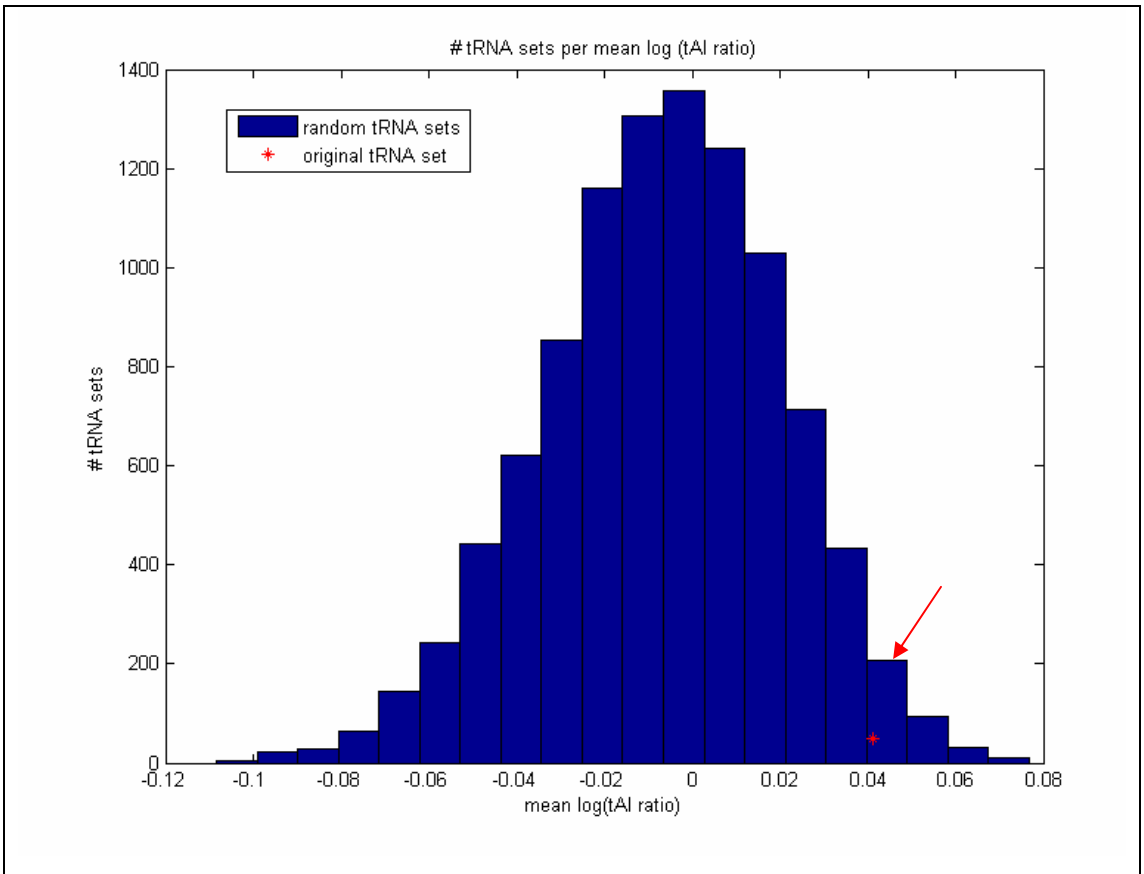


Figure 15: # tRNA sets per tAI ratio mean. A histogram of the number of random tRNA sets per log of the tAI ratio is shown. The ratio is calculated by dividing the tAI with the viral tRNA genes by the tAI without the viral tRNA genes for each viral gene, and the mean for each tRNA set is presented. The mean given by the original set of tRNA genes is marked in red

Next, I wanted to see if other tRNA sets could generate the separation we see between the effect on the viral genes and on the host genes. The separation was measured as the difference in the means of the populations normalized by the standard deviation:

$$\frac{\text{mean}(\Delta\text{virus\_tAI})-\text{mean}(\Delta\text{host\_tAI})}{[\text{std}^2(\Delta\text{virus\_tAI})+\text{std}^2(\Delta\text{host\_tAI})]^{1/2}}$$

where the ΔtAIs were calculated as the log ratio of the tAIs with the viral tRNA genes and without the viral tRNA genes. Figure 16 shows the histogram of the measured separation; the red arrow depicts the separation incurred by the actual gene set. It is clearly seen that the chosen set is one of the best available sets, with only 0.2% of the sets giving a separation measure which is higher. This indicates that the choice of the six viral tRNA is highly non-random and that the select set was optimized to enhance the translation of the viral genes on the expense of the host genes.

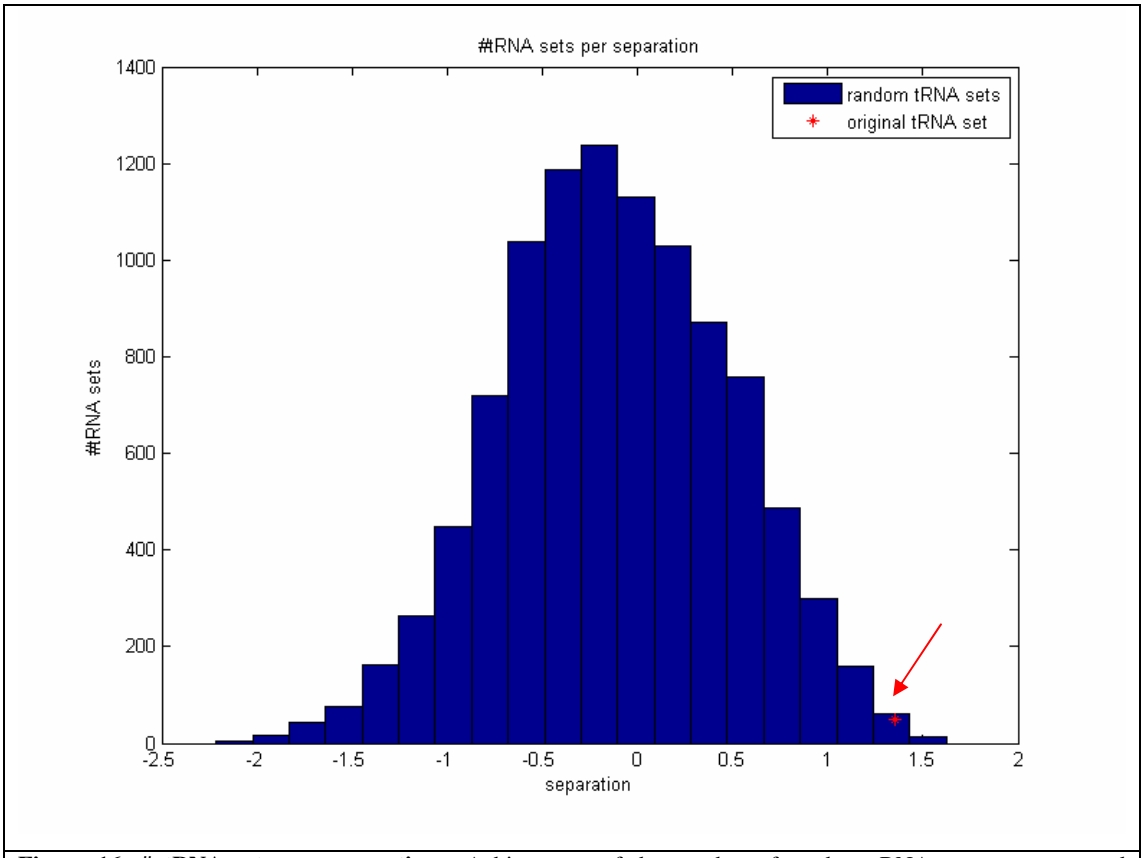


Figure 16: # tRNA sets per separation. A histogram of the number of random tRNA sets per computed separation is shown. The separation is measured as explained above. The separation given by the original set of tRNA genes is marked in red.

Testing the original set and the random set on *Synechococcus* WH7803 yielded a little less optimal results (top 3.5% and top 0.44% for the two measures respectively), while results on other hosts show significantly less optimal results, positioning the

chosen tRNA set as low as the bottom 10% in the *HL-Prochlorococcus*. These results indicate that the six tRNAs brought by the virus are highly optimized to work in the *Synechococcus* host background.

One might ask if six tRNAs is the optimized number of tRNA genes the virus should carry. Conceivably, less tRNA genes could give rise to the same effect on the viral genes population, but with a reduced cost of transcribing and coding for less tRNAs. On the other hand one can imagine that additional viral-encoded tRNA genes could create a stronger translation enhancement effect. To test this, I generated all possible tRNA sets which include an addition of a 7th tRNA gene, and all possible sub-sets of 5 tRNA genes derived from the original set of six, and tested the above scoring functions compared to the original six tRNA set. The results showed that none of the five tRNA sets gained a higher mean or separation compared to the original set, suggesting that five tRNAs are not enough to get the effect gained by carrying six genes. However, more than 20% of the seven tRNA sets generated a higher separation than the original set, and 50% of the seven tRNA sets generated a higher mean improvement compared to the original set. This suggests that other selective forces work in Syn9 against carrying seven tRNA genes or more.

2.3. Spatial patterns in translation efficiency

Since translation is a dynamic process, it is reasonable to look at the translation efficiency as a dynamic property of a protein coding sequence. For that purpose, I introduce here the translation efficiency profile of a sequence. Rather than looking at a single efficiency score for a gene, the translation efficiency profile is given by looking at an efficiency value for each position along the coding sequence. Thus, we might be able to detect patterns that govern the translation process of a coding sequence. Specifically the translation efficiency profile is given by looking at the tAI value of each position in the coding sequence. I chose to analyze the same eight yeast species that were used in the section 2.1, together with *E.coli* as a non-eukaryote representative.

The tAI value for each codon was calculated according to (dos Reis et al. 2004). The local profile of a gene was defined as the vector of the tAI values assigned to the gene's codons (omitting the first AUG), i.e.:

$$Local\_tAI_{Gene_i} = (tAI_{c_2}, tAI_{c_3}, ..., tAI_{c_n})$$

where c_i is the codon at position i in the gene (c_n is the codon before the stop codon).

For a particular species, all the genes in the genome were lined up once according to their start codon, and once according to their stop codon, and an average head and tail profiles were calculated as:

$$\overline{Local\_tAI_{start}} = (\overline{tAI_2}, \overline{tAI_3}, \overline{tAI_4}, ...)$$
$$\overline{Local\_tAI_{end}} = (\overline{tAI_n}, \overline{tAI_{n-1}}, \overline{tAI_{n-2}}, ...)$$

where:

$$\overline{tAI_i} = \sum_{Genes} tAI_{c_i} / |Genes|$$

2.3.1. The translation speed profile shows a conserved non-decreasing trend

The analysis showed a striking trend. In each species, the averaged profile starts with relatively low tAI values, which increase as the distance from the start increases. After the first ~50 codons, the profile reaches a plateau, and then starts to increase again at the last ~50 codons. The results can be interpreted as a tendency of the coding sequences to be translated slowly at the beginning of the sequence, with an

increase in speed as the ribosome moves along the sequence. In order to verify that this trend is a result of a specific and deliberate order of the codons in a gene, the mean and standard deviation of the profiles of 100 sets of randomized genes were calculated. In each randomization, the order of the codons was randomized for each gene independently, and the above calculation was performed on the randomized set. This procedure was repeated 100 times. The results clearly show that a randomized order of the codons does not show the trend of the true genome. Moreover, the true profile is different by more than three standard deviations from the randomized profile at the first and last ~50 codons, pointing to a very high statistical significance of the profile (figure 17).

In order to asses how much of the profile is governed by the amino acid (AA) sequence and how much by deliberately selecting low or high tAI codons, a new measure, the Expected tAI (EtAI), was introduced. In EtAI, all the codons of an AA have the same tAI value: a weighted average of the normal tAI values of all the AA codons. The weights are given by the background frequency of the codon usage for the given AA. EtAI implies that selection was acting only on the AA sequence, and not for a particular codon. Two genes with the same AA sequence, but different codon usage will have similar EtAI. Mathematically the measure is calculated as follows:

$$\mathrm{EtAI(AA)} = \sum_{Ci \in codons(AA)} codon\_usage(Ci) \times tAI(Ci)$$

$$codon\_usage(Ci) = \#Ci / \sum_{Cj \in codons(AA)} \#Cj$$

Figure 17 shows that the EtAI profile recapitulates only partially the actual tAI profile; especially, the low signal observed in the actual tAI profile at the beginning of the proteins is only weakly seen at the EtAI level. This indicates that the AA sequence can only partially explain the low profile at the start of the alignment, and much of it must be due to selection for low-tAI codons. On the other hand, the EtAI profile resembles more the actual tAI profile at the end of the protein suggesting that the actual signal of enhanced tAI at proteins' ends might be attributed partially to AA constraints.

In addition, I wanted to examine the effect of the codon usage on the observed profile. For that purpose, I also generated the averaged tAI for each AA, without considering the background information about the codon usage, i.e.,

$$tAI(AA) = \sum_{Ci \in codons(AA)} tAI(Ci) / \left| codons(AA) \right|$$

and calculated what the averaged profile will look like. This measure predicts what would be the translation profile if, given an AA sequence for a gene, the codons are selected from a uniform distribution. The most notable feature is that the profile is much lower (green line on figure 17). This means that there is a primary force that acts to elevate translation efficiency irrespective of sequence position. This is modulated by a secondary force that selects relatively slow codons at the beginning of the gene and high codons at the end of the gene. Also, it can be seen that the profile still has the trend of lower efficiency at the start and higher at the end, suggesting a selection for slowly translated amino acids at the start of a gene and fastly translated amino acids at the end of a gene.

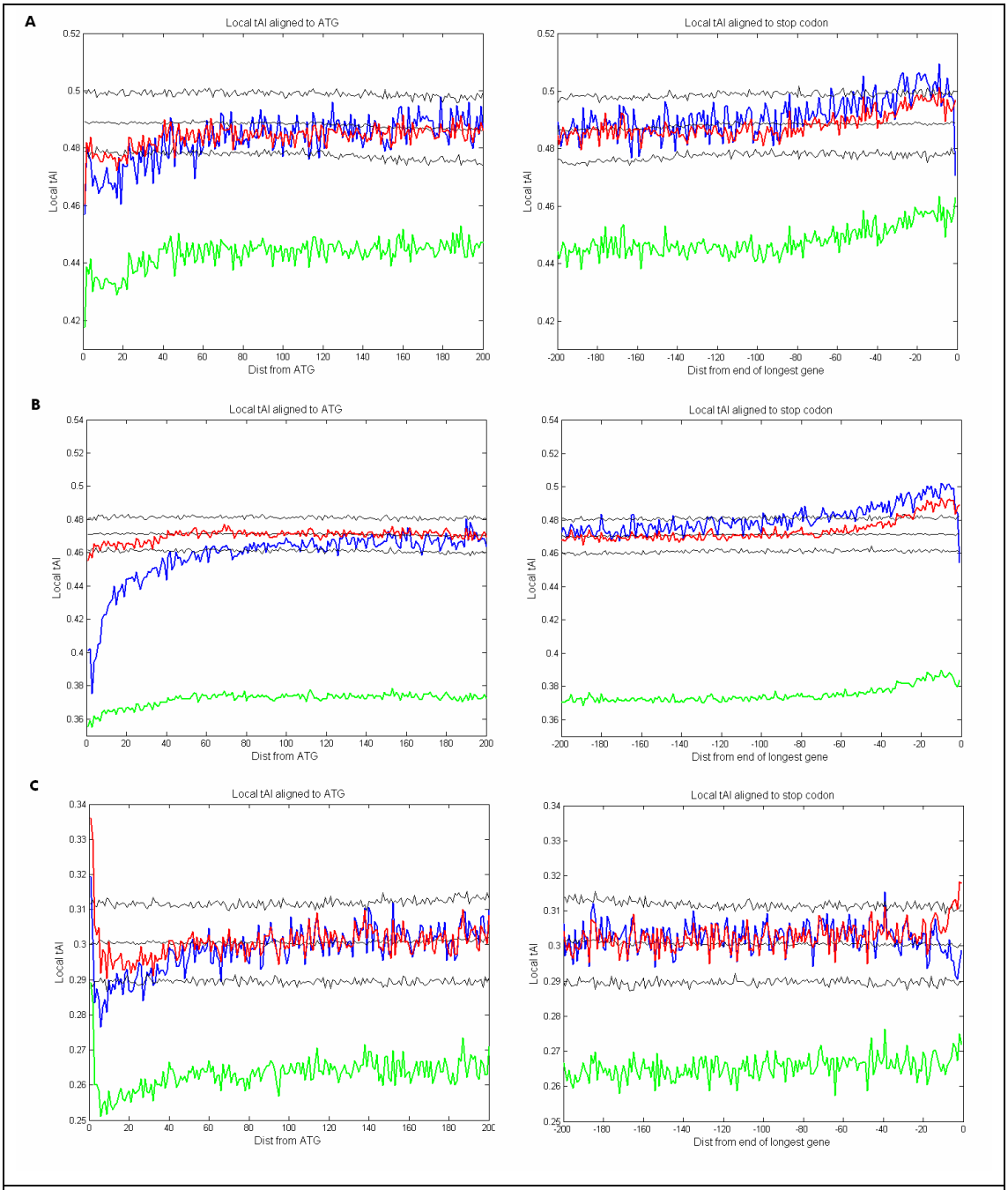


Figure 17. averaged tAI profile of 3 species. The first 200 codons are shown for the start codon line-up, and the last 200 codons are shown for the stop codon line-up.

A-*S.cerevisiae*, B-*Y.lipolytica*, C-*E.coli*.

each figure contains the averaged profile (blue), the expected profile(red), the AA averaged profile (green) and the randomized profile +-3 standard deviations (black).

The rest of the analyzed species can be found in appendix 4

One of the factors that can dictate a specific choice of codons is the GC content of codons. This factor might be especially influential in *Y.lipolytica*, where the GC content is relatively high (53%) compared to the rest of the analyzed hemiascomycotic species (37%-41%) (Dujon et al. 2004)). To test the hypothesis that the profile is generated due to a gradient in GC content along the protein sequences, I

created random coding sequences for *Y.lipolytica* and *S.cerevisiae* that preserve the original GC content at each position (see Methods). Figure 18 clearly shows that maintaining the per-position GC content profile of genes in *Y.lipolytica* is not sufficient for creating the original translation profile. The profile of the GC-preserving sequences, depicted by the blue curve, is relatively flat, indicating that the signal is lost when we only preserve this property. The same results are true to *S.cerevisiae* (data not shown).

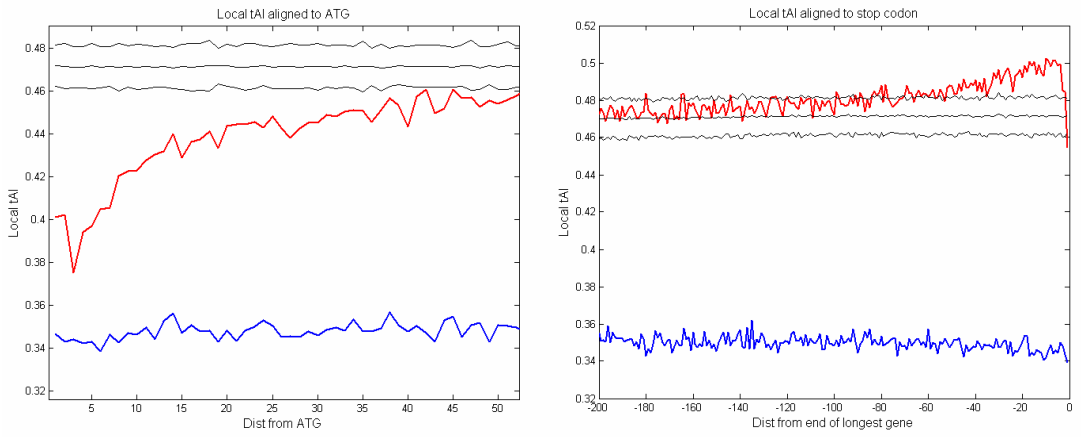


Figure 18: The effect of GC content on Local tAI of *Y. lipolytica*. Each figure contains the averaged tAI profile (red), the normalized profile +-3 standard deviations (black) and the averaged profile of the randomized genome (blue). The Genome was randomized with the constraint of preserving the GC content of each codon position. The profile clearly shows that keeping the GC content of a gene is not enough to generate the local tAI increasing profile.

Next, I turned to investigate the potential role of co-evolution between the tRNA pool and the ORF sequences of each species in the conservation of the observed pattern. I mainly aimed at distinguishing between two alternatives: one is that the translation efficiency profile is conserved because basic underlying features such as the tRNA pool and codon biases are conserved, with the alternative being that the profile is conserved despite the fact that both underlying features evolve. The latter would indicate selection for the observed profile that is a result of co-evolution of the tRNA pool and the coding sequences. For that I utilized the ability to generate computationally a "hybrid species" that inherits the tRNA pool from one species and the coding sequences of another species. Specifically, I performed an analysis of *S. cerevisiae* coding sequences, using *Y. lipolytica* tRNA pool, together with the reciprocal analysis. Figure 19 shows that the increasing trend of the profile at the start is significantly weakened when using the non-native tRNA pool (A-B,E-F, depicted

with a red arrow), suggesting that the coding sequences and the tRNA pool evolved together to maintain this trend despite variation in each of them. The results in the end show the same effect, but with a much weaker intensity (C,D,G,H).

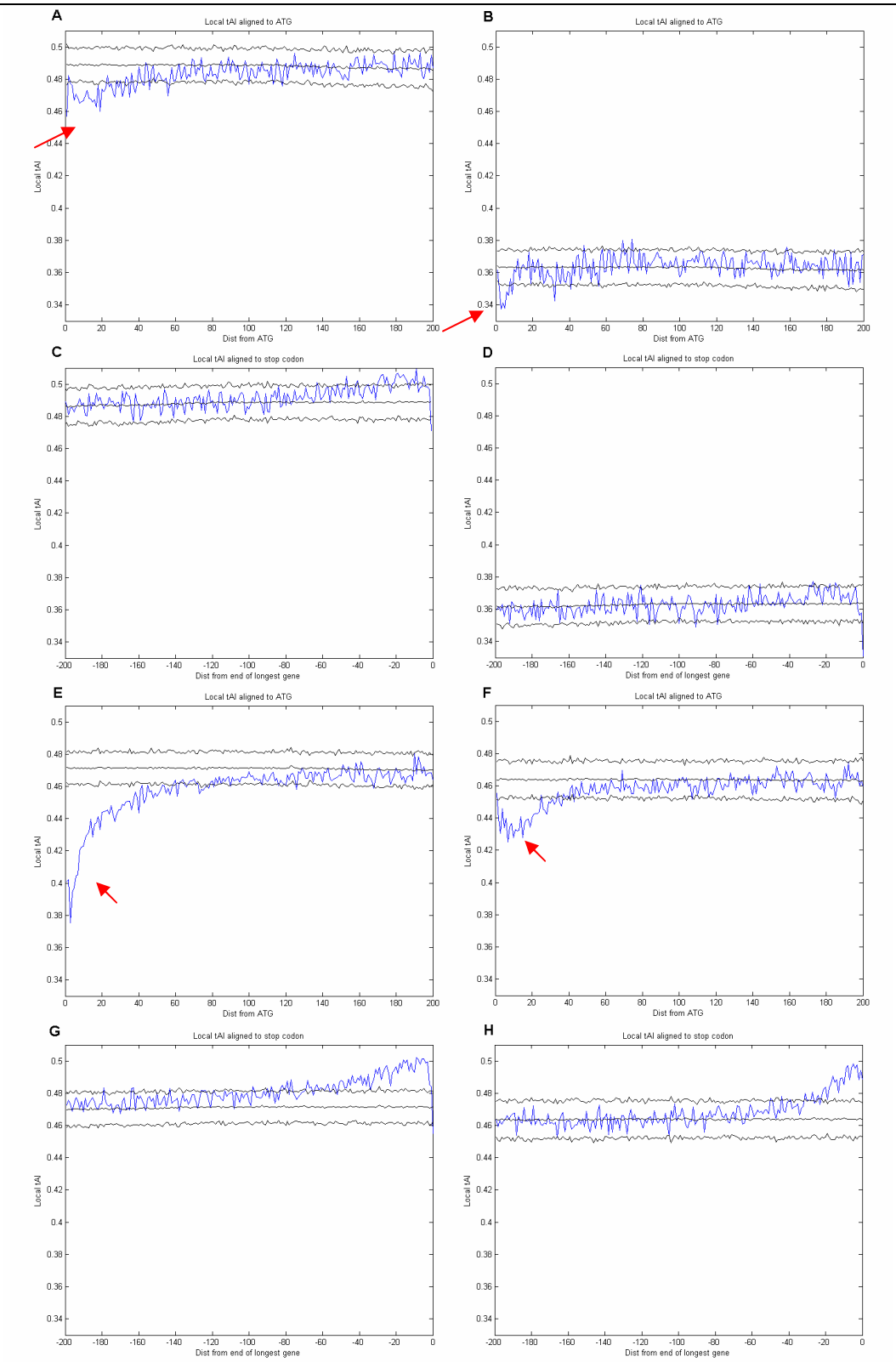


Figure 19: tAI profiles with native and non-native tRNA pools

A,C - the tAI profile of *S. cerevisiae*(start, end) B,D - the tAI profile of *S. cerevisiae* using *Y. lipolytica* tRNA pool(start, end).
E,G - the tAI profile of *Y. lipolytica*(start, end) F,H - the tAI profile of *Y. lipolytica* using *S. cerevisiae* tRNA pool(start, end)
the blue line represents the actual calculated tAI profile.
the black lines represent the mean +- 3 standard deviations of the tAI profiles of randomized sets of gene
The red arrows depict the weakened profiles. The left figure denotes the actual profile, and the right denotes the hybrid profile

2.3.2. The non-decreasing translation efficiency profile may be used to reduce ribosomal collisions

Since most transcripts are simultaneously translated by multiple ribosomes, over 10 ribosomes in some cases (Arava et al. 2003), this “translation speed profile” may be particularly crucial in minimizing “ribosomal traffic jams” on transcripts.

In order to test this hypothesis, I wrote a computer program to simulate the movement of ribosomes on an mRNA and record the number of collisions.

To simplify the model, several assumptions were made:

1. The rate of ribosome and tRNA binding depends only on the initial concentration (i.e., the number of ribosomes and tRNA currently attached to the mRNA does not affect the binding rate);

2.only three codon types exist: "Slow", "Medium" and "Fast", corresponding to codons with low, medium and high tAI, respectively; and,

3. when a ribosome collides with a preceding ribosome, the tail ribosome falls from the transcript and fails to complete translation.

The simulation was done on a 150 codons long mRNA, each run was simulated on a different sequence layout of the mRNA. The different layouts were generated by creating random permutations of 30 "slow" codons, 30 "fast codons" and 90 "medium" codons.

The translation profile results from the previous section indicate that the average gene strongly prefer the usage of low rate codons at the first 20-50 codons and to a lesser extent high rate codons at the last 20-50 codons. Accordingly, I defined an "ideal" mRNA sequence to contain the 30 slow codons at the start and the 30 fast codons at the end. The translation rate profile of every other sequence was measured in reference to the translation rate profile of this ideal sequence. The similarity of a sequence was determined by its correlation to the reference sequence.

This reference sequence is "ideal" in the sense that if we assume a deterministic movement rate, the speed of translation for this sequence is a non decreasing function of the ribosome position. Once a ribosome is attached to the mRNA, its speed will only increase as it moves to the end, and no collisions will be detected.

For each mRNA translation speed profile, a "success rate" of translation was calculated, which is the ratio between the amount of ribosomes that finished translation and the amount of starting ribosomes.

The simulation steps were performed as follows:

For each time step t:

1. Check if a new ribosome can bind to the mRNA (the recently bound ribosome is at least 13 codons ahead, since a ribosome cover approximately 10-15 codons).

2. Try to bind a ribosome with probability $W_b dt$, where W_b is the rate of ribosome binding.

3. If a new ribosome was bound – update its current codon movement rate according to the rate of the first codon.

4. For every other ribosome on the mRNA –try to move the ribosome with probability $W_{mx} dt$, where W_{mx} is the rate of movement for codon x.

5. If a moving ribosome attempts to move to an occupied location – release it and count it as failed. If a moving ribosome reached the end – remove it and count it is succeeded. Else – update the position of the ribosome and the movement rate.

The probability for movement / binding was calculated based on the rate with the following equation: $Pb = 1 - e^{-W*t}$ where w is the rate and t is the time step.

Figure 20 shows clearly that there is a significant correlation between the score of a layout and the success rate of the translation. The correlation coefficient is 0.727 with p-value < $2.2251*10^{-308}$.

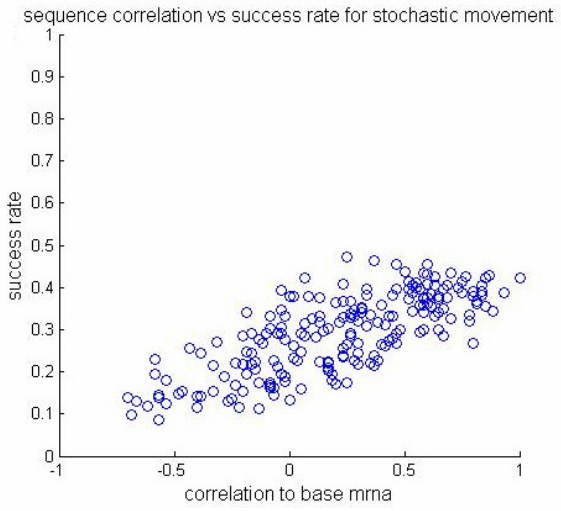


Figure 20: Correlation between mRNA order and translation success rate. Pearson correlation coefficient was calculated between the mRNA similarity score and the resulted success rate from the simulation. The correlation coefficient is 0.727, p value < $2.2251*10^{-308}$.

In order to test which region of the sequence has the most prominent effect on the success rate, I have determined the average rate of the translation from the start codon to every other codon in the sequence, resulting in 150 averaged rates for each mRNA. I also calculated the averaged rate from the end of the sequence to every other codon in the sequence. For each position, I then calculated the correlation between the averaged rates to that position and the success rate. Figure 21A, which shows the correlations between the success rate and averaged rates, calculated from the start codons, shows clearly that the rate of the first 13 codons is the most anti-correlated to the success rate, i.e. the slower the first 13 codons are, the higher the success rate ($r = -0.8134$, p-value $< 4.5925*10^{-237}$). Figure 21B, which shows the correlation between the success rate and the averaged rates calculated form the stop codon, shows that the faster the sequence from the 11^{th} codon onward, the higher the success rate ($r = 0.9165$, p-value $< 2.2251*10^{-308}$). Note that although the profile discovered in the previous section shows a tendency to increase towards the end of the sequence, the simulation does not show that the last codons have a significant role in determining the success rate. The scatter plots of the averaged rate against the success rate (Figure 21C, D) show the high correlation of the two measures.

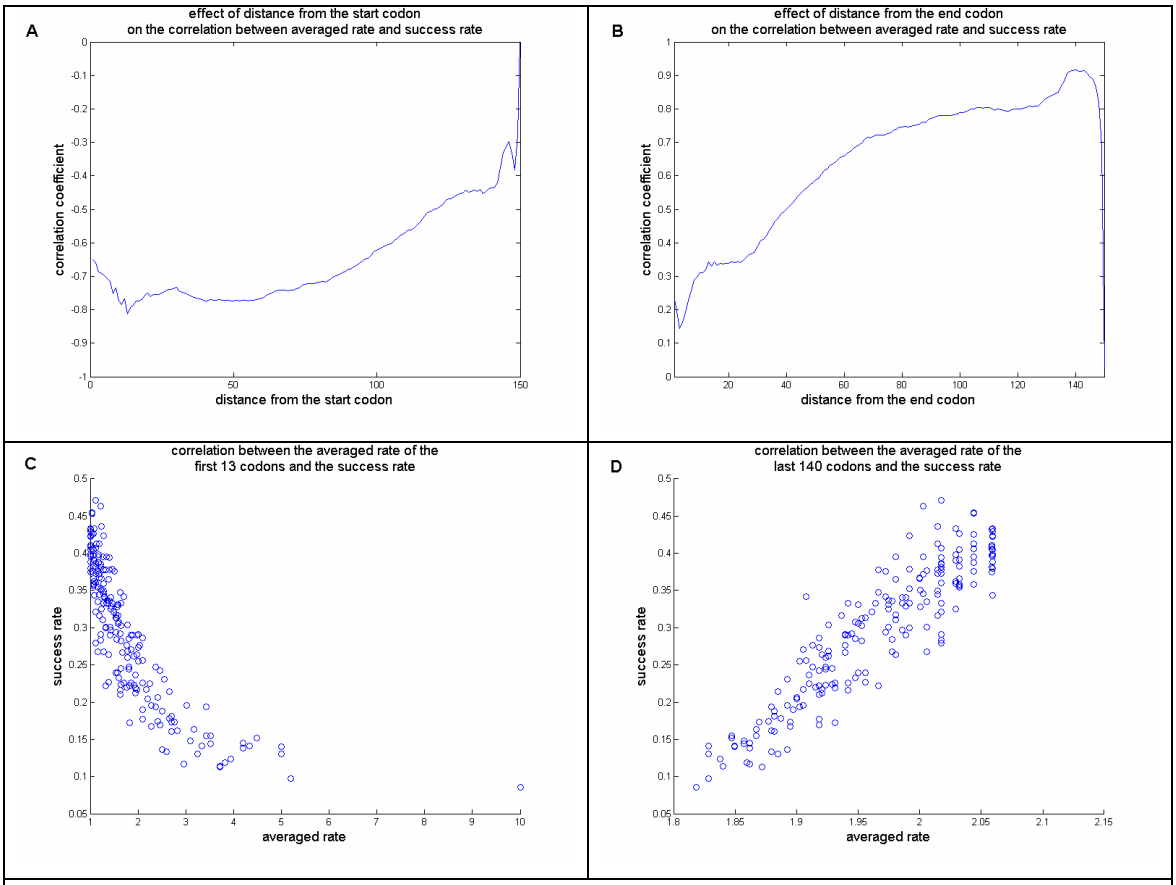


Figure 21: correlation between averaged translation rates and success rate.

A. correlation between the success rate and average rate (y axis) depending on the number of codons from the start codon to average(x axis)

B. A. correlation between the success rate and average rate (y axis) depending on the number of codons from the stop codon to average(x axis)

C. Scatter plot of the average rate of the first 13 codons and the success rate

D. Scatter plot of the average rate of the last 140 codons and the success rate

Figure 22 shows the correlation between the averaged translation rate and the success rate after running the simulation on ~4800 real *S. cerevisiae* coding sequences. The assignment of codons to "Slow", "Medium" and "Fast" was according to their tAI values in *S.* cerevisiae (see Methods). The results clearly show that the distinct effect of the rate of the first ~13 codons on the success rate is apparent also in real sequences. This effect sharply disappears after the 13th codon (Figure 22A). No effect is found near the stop codon (Figure 22B).

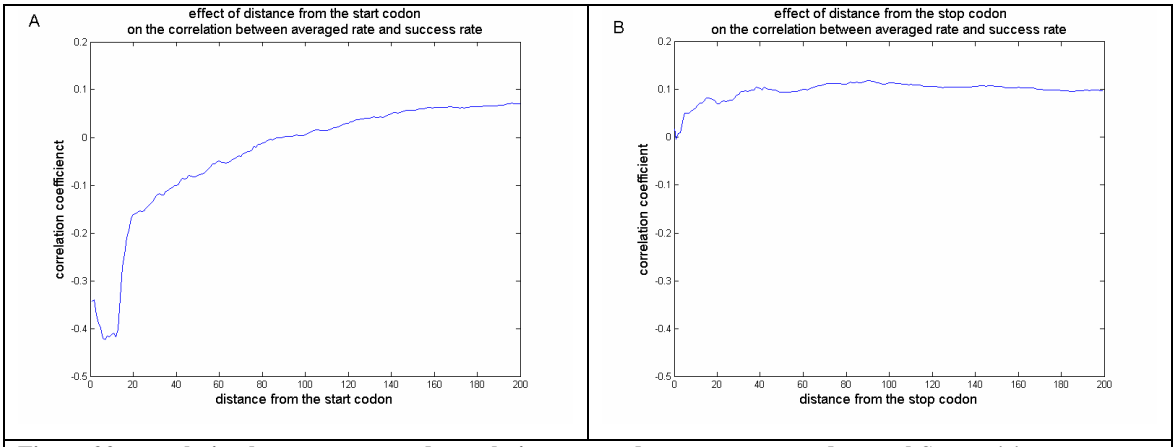


Figure 22: correlation between averaged translation rate and success rate tested on real *S. cerevisiae* sequences.

A. correlation between the success rate and average rate (y axis) depending on the number of codons from the start codon to average(x axis)

B. A. correlation between the success rate and average rate (y axis) depending on the number of codons from the stop codon to average(x axis)

Since according to the simulation parameters the ribosomes have to be at least 13 codons apart from each other, it is no surprise that a cluster of slow codons about this size at the beginning will be enough to reduce the probability of collision between two ribosomes. Once a ribosome clears the bottleneck at the start of the sequence, a newly attached ribosome will be held there for a while, giving its preceding ribosome enough time to accumulate distance from the site and avoid collision. I thus expect that different parameters of ribosome sizes or codon rates may change this value.

3. Discussion

The three studies presented here have shown how the tRNA pool and the coding sequence have co-evolved in order to regulate the translation efficiency of proteins. In the first study, I have shown using a computational tool which I developed, that differences in the translation efficiency among organisms are generated by changes in the coding sequences, changes in the tRNA pools, and the interplay between them. In the second study, I have shown how a virus changes the tRNA pool of its host in order to compensate for the differences in their codon usage. I have shown that this adaptation is highly optimized to a specific host genomic background. In the third study, I have shown the existence of a local translation efficiency pattern which governs the translation efficiency in many organisms, and how both the coding sequence and the tRNA pool are needed to co-evolve in order to conserve this pattern in different organisms.

Throughout this study, I have used the tAI (dos Reis et al. 2004) for measuring translation efficiency. The choice to use the tAI was considered best for several reasons: First, it is assumed to measure translation efficiency more directly, rather than measuring the codon bias, which previous studies have measured. Second, unlike other measures, no prior knowledge about the highly expressed genes in the genome is needed in order to apply the tAI for a sequenced genome. This allows us to systematically test the translation efficiency profiles on a large number of organisms which were not examined elsewhere. Third, the tAI, by incorporating both the tRNA availability and the coding sequence adaptation to it, can be used to examine which of these two factors contributes to differences in the observed translation efficiency. And last, its codon based calculation, gives the ability to generate with it a local translation efficiency profile. Although other indexes, such as the CAI, share this ability, they do not share the aforementioned advantages that the tAI bear.

3.1. *Cis* and *Trans* changes affecting translation efficiency

In this study, I have developed a computational method to differentiate between *cis* and *trans* changes which lead to differences in translation efficiency among orthologs genes. *Cis* and *trans* are concepts borrowed from the transcription context, in which *cis* denotes factors that have a local, short range effect on genes, while *trans* denotes

factors with a wider effect, usually affecting groups of genes [see, for example, (Wittkopp et al. 2004)]. In this study I applied the above-mentioned concepts in the context of translation: changes to translation efficiency caused by changes in coding sequence were considered *cis* changes, while changes caused by modifications to the tRNA pool were considered *trans* changes.

The translation efficiency difference between genes is calculated as the ratio between the two genes tAI. The ratio was chosen, as the tAI is a relative measure, thus differences in the tAI are essentially differences in ratios. The tAI is a relative measure since its calculation involved normalizing to a reference point and thus it is measured in arbitrary units. The original tAI is calculated by normalizing each codon weight to the maximum weight, such that the most efficient codon gets a tAI of 1. With this normalization, the gene that would get a tAI of 1 is a gene with the maximal efficiency, i.e. a gene that uses only highly efficient codons. In this case, all the genes are measured in reference to this gene. However, this normalization is very sensitive to changes in the maximal number of tAI genes. While this is a little concern when examining the tAI of an individual organism, this becomes very critical when comparing the tAI between two species, as it can create differences that arise solely due to small changes in the maximal number of tRNA gene copy numbers. To avoid this problem, the tAI was normalized in this study to the averaged genome tAI (see Methods). With this normalization, the gene with a tAI of 1 is a gene that has the exact codon frequencies as the entire genome. All the other genes are measured in reference to this gene, and thus a gene can get a tAI which is smaller or greater than 1. Note however, that as a result of the normalization in the computation, changes in *trans* do not account only for differences in the tRNA pool, but also to the differences in the overall codon bias between two organisms, as these differences change the tAI. Nevertheless, global changes to codon bias may still be considered as *trans* changes, as they affect the translation efficiency of the entire genome.

The new normalization is more robust, since the average is less sensitive to small changes in the tRNA gene copy numbers. In addition, the current normalization gives another advantage. The original tAI takes into account only the "supply" of tRNA molecules, as it considers the amount of tRNA molecules as the determinant of the translation efficiency. However, it is obvious that the amount of tRNA molecules must meet the demand for them. If many genes in a genome require the abundant

tRNA molecules, they may be less efficient compared to a background were only a few genes require the abundant tRNA molecules. In the latter case tRNA molecules will be more available when needed. By introducing the codon usage to the score and thus taking into account the amount of codons that correspond to each of tRNA molecules, in addition to increasing the robustness of the measure, I was able to incorporate some form of the demand for the tRNAs as a factor in the score. These normalization factors correspond to the static demand and supply of the tRNA pool and may be generalized to cases in which both the mRNA and tRNA pools are dynamic. However, at present, there are not sufficient data to perform such generalization.

I chose to analyze eight yeast species. These organisms were chosen since they are close enough in order to have significant amounts of ortholog genes, but remote enough to show diverse coding sequence similarities and tRNA pool similarities.

The results show that while all components contribute to the translation efficiency difference between genes, *cis* is the major contributor both to the changes between organism pairs and to changes in ortholog sets within pairs. These results are in accordance with the observation by Man & Pilpel (2007), that the tRNA pools of the analyzed species evolved slowly, thus changes in *trans* are expected to be weak. Nevertheless, I was able to detect genes in which changes in *trans* were the major contributor to the translation efficiency difference, even in species with highly correlated tRNA pools, such as *S. cerevisiae* and *K. lactis*. A particularly interesting example was obtained with the DNA repair enzymes in which their high translation efficiency in *S. cerevisiae* can now be attributed to changes in the tRNA pool. Such genes may serve as good candidates for further study, since they may hint about environmental stresses that selected for changes in the tRNA pool. For instance, an interesting possibility is that the natural habitat of *S. cerevisiae,* on grapes exposed to the sun, may have contributed to the shaping of its tRNA pool in a way that translation of DNA repair enzymes would be enhanced.

One of the most interesting species in the analysis is *Y. lipolytica*. This organism shows the most distinctive tRNA pool compared to other organisms analyzed in this study. For example, its correlation to *S.* cerevisiae tRNA pool is only 0.58 (Man and Pilpel 2007). This observation is reflected in the calculations of tAI decomposition, as the *trans* component showed a unique significant effect on the tAI ratio in pairs

involving *Y. lipolytica.* All pairs involving *Y. lipolytica* showed a biased *trans* component, with lower translation efficiency in *Y. lipolytica*, and the *trans* component is showed to be the most correlated to the tAI change in those pairs. An interesting question is what caused the shift in the tRNA pool. One can consider two alternatives that resulted in a change in the tRNA pool and the codon bias. The first is the bottom up approach, in which the codon bias was under a selective pressure, and the need to change the codon bias resulted in an adaptation of the tRNA pool to the change. The second alternative is the top down approach, in which a selective pressure on the tRNA pool resulted in an adaptation of the codon bias to correspond to the change. Since *Y. lipolytica* has a unique GC content usage compared to other hemiascomycotic species (53% in coding sequences, compared to 40% in *S. cerevisiae*), (Dujon et al. 2004), it might serve as a driver for the change. Indeed, examination of the correlation between GC content of the tRNA anti-codon and tRNA gene copy numbers (tGCNs) showed that while GC content and tGCNs are anti-correlated in *S. cerevisiae* ($r=-0.4$, p-value<0.0058, spearman's one sided rank correlation), indicating that high copy number tRNA genes tend to have low GC content, this anti-correlation is lost in *Y. lipolytica* ($r=0.06$, p-value<0.7, spearman's rank correlation). In addition, the tGCN difference between codons, taken as the ratio between *S. cerevisiae* codons tAI and *Y. lipolytica* codons tAI, also showed a weak anti-correlation to the GC content ($r=0.34$, p-value<0.0036, spearman's one sided rank correlation). This finding supports the bottom up possibility, namely that GC content was increased in the YL genome, and this affected codon bias, which in turn, affected tRNA GCNs.

The method developed here was used to analyze yeast species, but it is not limited to the current choice of organisms. Since it relies solely on computational procedures, it can be easily applied to any pair of organisms in which we can derive the tRNA gene copy numbers, overall codon bias and find orthologous genes. Specifically, any organism with a completely sequenced genome can be analyzed by this method, provided that codon bias in the analyzed species is shown to be governed by translation efficiency (dos Reis et al. 2004).

3.2. The effect of viral tRNA genes on the translation efficiency of viruses and their hosts

In this study I have shown how tRNA genes carried by a virus affect the translation efficiency of the virus genome and the host genome upon infecting the host. This effect serves as a unique and interesting case of *trans* change, since the virus changes the cellular tRNA pool of the host it encounters. I have chosen to analyze *cyanophage* Syn9, a bacteriophage infecting marine *cyanobacteria*. This phage carries in its genome six tRNA genes, and was shown to infect a wide range of bacteria under lab conditions (Sullivan et al. 2003).

I have shown that the six tRNAs encoded by the virus are highly optimized to elevate the translation efficiency of its genes under the genomic background of bacteria from the *Synechococcus* genera, its native host. In the background of most of the potential host bacteria from the *Prochlorococcus* genera, the tRNA from the virus seemed to decrease the translation efficiency of its genome. The hosts are "potential" since infection of these bacteria was tested only under lab conditions, and no evidence was yet found regarding the infection under natural conditions. The decrease in the translation efficiency of the viral genome happened regardless of the infection susceptibility of the potential host, i.e. even in hosts which were shown to be susceptible to infection by the virus, its genome translation efficiency was decreased.

At this point, one might ask how an addition of tRNA genes can decrease the translation efficiency of some genes. While the cause for an improvement in the translation efficiency is straightforward, given that the gene's codon might have more tRNA molecules that correspond to them, the decrease is a little less intuitive. In practice, the decrease comes to effect by normalizing the codons' tAI as explained in the Methods section. An addition of tRNA genes will increase the normalization factor, thus decreasing the tAI of codons which do not correspond to the added gene. Biologically, under the assumption that cell resources are fixed, a decrease in the translation efficiency makes sense, if a gene does not have many codons that correspond to the added tRNA. By adding a tRNA gene that needs to be transcribed, the transcription resources have to be divided between more tRNA genes, thus reducing the resources allocated to all other tRNA genes. This might cause a reduction in the number of tRNA molecules that are transcribed from tRNA genes which were not added a copy, thus reducing the translation efficiency of genes that

mostly contain codons which corresponds to those genes. Another consideration, regardless of the transcription resources, is the probability to encounter a wrong tRNA. Adding more tRNA genes to the cell shifts the relative concentration of tRNA molecules towards the tRNAs which are encoded by the added genes, and thus, excluding the codons which correspond to the added molecules anti-codons, there is a higher probability for a codon to encounter a non-cognate tRNA. This will decrease the translation efficiency of the codons, and depending on the ratio between the elevated codons and the lowered codons, the translation efficiency of a gene might be improved or worsened.

The observation that the tRNA genes encoded by the virus decrease the translation efficiency of its genome in potential hosts is a little bit puzzling. It seems unreasonable for the virus to keep tRNA genes that would impair its ability to infect its hosts. However, the codon usage of the *Prochlorococcus* genera is more similar to that of the phage, and examination of the translation efficiency of the viral gene showed that even after the decrease caused by the viral tRNA genes, the averaged tAI is still higher in those bacteria than in the *Synechococcus* (data not shown). These results suggest that by carrying the tRNA genes, the Syn9 virus was able to expand its host range, supporting previous hypothesis about the *Chlorella* virus by Nishida (Nishida et al. 1999), which suggested that tRNA genes are needed in order for a virus to adapt to a wide range of hosts with different codon usage. More specifically, the tRNA genes may be used at a status of “break in case of emergency”, namely they are not needed with the usual hosts – for these changes in the coding sequence may have already optimized the viral genes. The tRNA genes may actually facilitate infection in relatively new hosts, to which the virus is not yet adapted.

I have used functional annotation of the viral genes, derived from (Peter R. Weigele 2007) to show the effect of the viral tRNA on different functional groups, and have shown that the genes most affected by the viral tRNA genes are the virion proteins, which are part of the “late" genes. Those genes are transcribed and expressed in the late stages of the viral infection, and are required in large amount, as they are part of the virus particle. These results coincide with results reported by Kunisawa (Kunisawa 1992), who showed that in the *coliphage* T4 the codon usage of late structural proteins is highly correlated to the tRNA genes encoded by the virus. It was also shown that the transcription of tRNA genes in T4 happens mostly from the “late”

promoters (Broida and Abelson 1985), which are promoters typical to genes transcribed late in the infection cycle. The observation that the tRNA genes are transcribed from "late" promoters coincides with the increased effect of the viral tRNA genes on "late" genes. There is lack of data about the Syn9 promoters or gene expression, but the strong effect of the tRNA genes on late Syn9 genes suggests that the tRNA genes of Syn9 will show similar behavior. Another observation made by Kunisawa is a strong anti-correlation between the abundance of proteins in the virion, and the codon bias towards codons corresponding to the viral tRNA genes. It was suggested that extremely abundant proteins, like the major capsid protein, which appears about 1000 times in each particle, would benefit more from the usage of the host tRNA molecules, which are assumed to be more abundant, leaving the virus tRNA molecule for the less abundant virion proteins. If the highly abundant proteins would need to use the viral tRNA molecules, which are assumed to be rare, they would deprive them from the lowly expressed proteins. This observation is true also in the Syn9 proteins. There is a significant anti-correlation between the improvement in the tAI of the virion proteins and their copy numbers in the virion ($r=-0.72$, p-value < 0.02), namely, the genes which are improved the most are the less abundant from the virion proteins. However, all the genes show an improvement in their tAI upon introducing the viral tRNA genes, suggesting that they all benefit form the viral tRNA. Note that the tAI is valid under the assumption that the concentration of tRNA molecules is proportional to their gene copy number. This assumption might not be valid when comparing viral and bacterial tRNA genes, as suggested by Kunisawa, as they might be transcribed on different time scales and from different promoters. However, due to lack of data regarding gene expression in the bacteria and the virus, we are unable to validate or reject either of the assumptions.

The T4 and Syn9 have shown to exert the same effect on their hosts' genomes, i.e., the elevation in the translation efficiency of the virion proteins. It is interesting to note that to generate the same effect they carry with them different tRNA genes. While the T4 phage carry with it eight tRNA genes (Kunisawa 1992), only three are shared with Syn9. Testing the effect of the entire set of T4 tRNA genes, and the effect of only the shared genes on the Syn9 genome resulted in less improvement in the translation efficiency, sometimes even a decrease (data no shown). These results also

demonstrate the need of the tRNA pool to co-evolve with the codon usage of the phage and hosts in order to generate similar effects in different environments.

Lastly in this study, I have shown the effect of the viral tRNA gene on the host genome. I have shown that a group of host genes also show an increase in the predicted translation efficiency as a result of introduction of the viral tRNA genes to the host system. While this group is mainly comprised of hypothetical proteins, some of the proteins are assigned a function, and those proteins are enriched for cell envelope and transport proteins. It was suggested previously that these envelope-modifying proteins might be used to assist cells in evading grazers and phages by changing the characteristics of the cell envelope (Palenik et al. 2003),(Monger et al. 1999). Since in the Syn9 infection, the translation efficiency of these proteins is enhanced after the viral infection, an interesting hypothesis is that these proteins may have a role in "superinfection exclusion", a known phenomenon in which phage-infected bacteria become immune to recurrent infections (Lu and Henning 1994). It is possible, that by enhancing cell envelope modifying proteins, the phage changes the cell envelope properties to prevent from other phages to attack the bacteria.

The ability of the viral tRNA genes to affect the host genes is debatable. It is known that sometime after infection, the host transcription and translation machinery is completely devoted to the transcription and translation of the viral genome, and thus, it is not certain if the viral tRNAs have the opportunity to affect the translation of bacterial proteins. No data exist on the time scale of the Syn9 infection cycle. However, a recent study was performed on a related *cyanophage*, S-PM2, which infects *Synechococcus* WH7803. The S-PM2 encodes 239 proteins (Mann et al. 2005) and thus is roughly the same size of Syn9, therefore it is reasonable to assume similar replication times for these two genomes. It was shown that the late genes of the virus are at peak expression 6 hours after infection, while the host transcripts, although in constant decline, are present in the cell for at least 10 hours after infection (Martha R. J. Clokie 2006). Those time scales, assuming they are similar in the Syn9 infection cycle, allow enough time for the viral tRNA to influence the translation efficiency of the host proteins, and thus making the above analysis relevant.

3.3. Spatial patterns in translation efficiency

By creating a translation efficiency profile for each gene based on the tAI, I have shown in this study the existence of a conserved trend in unicellular prokaryotes and

eukaryotes genomes. This trend suggests that on average, the translation efficiency increases as the distance from the start codon increases. More precisely, there is a gradual increase in the translation efficiency in the first ~50 codons, followed by a plateau, and then a gradual increase again in the last ~50 codons. I have also shown that this trend can only be partially explained by a selective usage of amino acids that could mainly be encoded by rare tRNAs at the start of a sequence and amino acids that mainly correspond to high abundance tRNAs at the end of a sequence. In contrast, I found that the observed trend can be mainly attributed to the translation efficiency level itself, with two forces working on this level: One force work towards globally elevating the translation efficiency of the coding sequences, and another, generating the positional modulation which account for less selection pressure on the start of a sequence.

Using the computational ability to create "hybrid species", that inherits the tRNA pool from one species and the coding sequences of another species, I have shown that the conservation of this profile required the co-evolution of the tRNA pool and the coding sequences. The observation that despite changes in the coding sequences and the tRNA pool, this trend was conserved, suggests a selective advantage for keeping this profile.

Several studies had already shown this trend in the past; see for example (Eyre-Walker and Bulmer 1993),(Bulmer 1988),(Chen and Inouye 1990). However, most of them tested only a sample of genes, and were not conducted on a genome-wide scale or on a wide range of organisms. Also, the translation efficiency in those studies was not measured directly, but using the codon bias as an indirect measure for translation efficiency. A more recent study conducted by Qin et. al. (Qin et al. 2004) tested the spatial pattern in codon usage on four prokaryotic genomes and two eukaryotic genomes. They have used an improved version of the effective codon number measure, Nc, (Wright 1990), (Novembre 2002) to test for patterns in the codon usage bias. They found an incremental pattern in the codon usage bias, i.e., there is less bias in regions close to the start codon, with increasing bias as the distance from the start codon increases. These findings are consistent with the translation efficiency pattern observed in the current study, as with no selection on the codon bias, the chance of lowly efficient codons to appear increases. Qin et. al. also found this pattern to be stronger in highly expressed genes, which would suggest that a similar pattern will be

reflected in the translation efficiency; however, this was not tested in the current study.

In the last part of this study I have tested a possible theory regarding the benefit a non decreasing translation efficiency pattern may bring. Using a simulation approach, I have shown that a non-decreasing profile, similar to that seen in biological sequences, can be used to decrease the number of collisions between ribosomes translating a single transcript, thus preventing ribosomal "traffic jams". The results of the simulation clearly show very high correlation between the success rate and the "ideality" of the sequence, measured by its similarity to sequence with a non-decreasing efficiency profile.

The most popular reasoning for the observed pattern is the increase in cost of a nonsense error as the elongation process advance. A nonsense error at the start of a transcript will waste less resources than a nonsense error at the end of the transcript, thus, the selection power is thought to increase as the distance from the start codon increases (Kurland 1992), (Akashi 2001), (Qin et al. 2004). While this hypothesis states the advantage of having an increasing profile over a relatively low profile, it does not explain why the observed profile is better than an overall highly efficient profile. Such a profile will reduce the chance of nonsense errors all over the sequence. One alternative is that this profile is used for expression regulation. Positioning codons which correspond to rare tRNAs near the start codon was shown to significantly reduce the expression of genes. By controlling the concentration of rare tRNAs, the expression of those genes can be regulated (Chen and Inouye 1990),(Chen and Inouye 1994). An alternative hypothesis is that the slow translation efficiency near the start codon serves to regulate and control the number of ribosomes on a single mRNA message. This would serve two purposes. First, it will reduce the number of ribosomes on the transcript, as was shown by Zhang *et al.* (Zhang et al. 1994). Assuming ribosomes are found in limited amounts in a cell, this would allow more transcripts to use the available resources, rather than letting many ribosomes to occupy a single mRNA. Second, as was shown by the current study, this will prevent traffic jams, which may cause ribosomes to generate queues, reducing their effective use. Note that contrary to the simulation, collisions in the translation process are not a rare event and usually they do not end in the failing of a ribosome to complete the translation. However, collisions do slow the rate of translation and increase the chance

of a ribosome to fail, thus for the simplicity of the simulation, the translation efficiency was reduced by "wasting" ribosomes after a collision. A recent study published during the course of my work addressed the issue of ribosomal collisions (Mitarai et al. 2008). The authors showed, using experimental and computer simulations, that collision are unavoidable, and that two strategies may be used to reduce them. The first is selection for slowly translated codons close to the start codons, a result which support the finding of the current study. The second strategy is to make the mRNA unstable, such that it degrades after a certain period of time, before ribosomes become highly stacked.

Finally, I would like to point that an interesting follow up to the studies presented here would be the conjunction of the two computational methods presented in this study, the tAI decomposition and the local tAI profile, to study the co-evolution of the tRNA pool and the coding sequences. The tAI decomposition method presented in the first chapter allows us to study the co-evolution of the coding sequence and the tRNA pool without explicitly addressing the order of the codons in the sequence. This approach was based on the observation that what influence the overall efficiency of a gene in terms of the tAI are its overall codon usage and its adaptation to the tRNA pool, and not the actual spatial arrangement of slow and fast codons along transcripts. However, this observation is a simplifying one, and it is conceivable that the mechanism in which evolution is working is local, i.e. changes in the coding sequences occur on a per base resolution, and not on a global gene scale. Thus in order to change the translation efficiency of the *cis* and co-evolution components, the gene would have to undergo a series of point mutations. Comparing the local profiles of orthologous genes with regards to their *cis*, *trans* and co-evolution changes will allow us to study the mechanism in which evolution generated the translation efficiency differences between the genes.

4. Methods

4.1. Methods used for the *Cis* and *Trans* changes affecting translation efficiency

4.1.1. tRNA gene copy numbers

tRNA gene copy numbers were taken from Man & Pilpel (2007), http://longitude.weizmann.ac.il/pub/papers/Man2007_tai/suppl/onlineSuppTables/tableS1.xls. For all species except *C. albicans* the tRNA gene copy numbers were obtained by applying the tRNAscan-SE software version 1.1 (Lowe and Eddy 1997), which uses a hidden Markov model (HMM)-based approach, to the genome sequences. For *C. albicans* the tRNA gene counts were extracted from the *Candida* Genome Database (CGD) (Arnaud et al. 2005). For *A. nidulans* the genome sequence was obtained from

http://www.broad.mit.edu/annotation/genome/aspergillus_nidulans; chromosome sequences for the remaining seven species were obtained from GenBank (http://www.ncbi.nlm.nih.gov/Genbank).

4.1.2. Protein and coding sequences

Protein and coding sequences for all yeast species analyzed were downloaded from Man & Pilpel (2007),

http://longitude.weizmann.ac.il/pub/papers/Man2007_tai/suppl/codingSequences/codingSequences.html)

The sequences on the site were obtained from several sources. The *C. albicans* protein and coding sequences were downloaded from http://candida.bri.nrc.ca (Braun et al. 2005). *S. cerevisiae* protein and coding sequences were downloaded from the Saccharomyces Genome Database (SGD) (Balakrishnan et al.). *A. nidulans* Protein, gene sequences and gene structures were downloaded from http://www.broad.mit.edu/annotation/genome/aspergillus_nidulans (Galagan et al. 2005).

For the remaining five species (*S. pombe, Y. lipolytica, K. lactis*, *C. glabrata*, *D. hansenii*), data was downloaded from Integr8 (Pruess et al. 2005) and the EMBL database (Kanz et al. 2005). Some of the coding sequences for *S. pombe* were

obtained *S. pombe* section of GeneDB (Hertz-Fowler et al. 2004). See (Man and Pilpel 2007) for technical details.

4.1.3. Normalization of the tRNA adaptation index (tAI) for coding sequences

The tRNA adaptation index is described in detail in (dos Reis et al. 2004, see Appendix 5). 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 (including wobble interactions). For a given coding sequence, the tAI value is then the geometric mean of the weights of all its sense codons. Originally, the weight for each codon is normalized by the maximal weight, resulting in the tAI of a coding sequence to range from 0 to 1. This makes the index highly sensitive to the maximum number of tRNA genes. To increase the robustness of the index, I normalized each codon to the averaged genome tAI. The averaged genome tAI was calculated by creating an artificial sequence, with codon usage frequency that correspond to the genome codon usage and calculating its unnormalized tAI. Normalizing each codon with this normalizing factor creates for each organism a tAI distribution which is centered on 1, where highly efficient genes receive tAI which is greater than 1, and lowly efficient genes receive tAI which is smaller than 1. The calculation was further modified to include the first codon, as well as other methionines.

4.1.4. Decomposition of the tAI to cis, trans and co-evolution components

4.1.4.1. Generation of a table of orthologous gene groups

Prior to decomposing the tAI difference into their underlying components accurate clustering of orthologs in all species was needed. Using the inparanoid algorithm (Remm et al. 2001), a two-species ortholog lists were constructed for every pair of species, using *C. neoformans*, a basidiomycotic fungus, as an outgroup. I used a modified version of the inparanoid program, supplied by Orna Man (see Man and Pilpel 2007). There is a discrepancy between the inparanoid algorithm, as reported by Remm et al. (Remm et al. 2001), and the programs 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 for the sequences to be considered homologous, the program applies this cutoff to the shorter

sequence. In order to avoid domain-level matches, the inparanoid program was modified to reflect the algorithm as presented in the paper. The MultiParanoid program (Alexeyenko et al. 2006) was used to merge these two-species ortholog lists into one matrix. The order of species in the input to the program was as follows: *A. gosypii, S. pombe*, *S. cerevisiae*, *A. nidulans*, *Y. lipolytica*, *C. albicans*, *K. lactis*, *C. glabrata*, *D. hansenii*, and *S. bayanus* (*A. gosypii* and *S. bayanus* were eventually excluded from the analysis) . With this order, two-species ortholog lists with large evolutionary distances between the relevant species, such as *S. pombe* and *A. nidulans*, were processed before ortholog lists of close species, such as *S. cerevisiae* and *C. glabrata*. The output of the MultiParanoid program was converted into a matrix of orthologs where each row corresponds to a gene and each column to a species. Note that if duplication had occurred after the divergence of *S. pombe* and *A. nidulans* from the remaining species, there would be more than one gene representing the same species in the same orthologous group (row). It was assumed that all genes in a single orthologous group have the same function, henceforth they are referred to as a single gene.

4.1.4.2. Computing the decomposition of the tAI

The tAI decomposition was calculated for each pair of organisms as described in the results (section 2.1.1). For each pair, the decomposition was computed for every set of orthologous sequences. If there was more than one gene from the same organism in a set, the gene with the highest tAI value was chosen for the analysis. Coding sequences that included stop codons, or did not start with ATG were excluded from the analysis.

4.1.5. Generating the random orthologous sets

(Section 2.1.2, figure 3 in Results) Each of the 1,000,000 sets of orthologs was generated by randomly choosing two tRNA sets and two coding sequences. To create a tRNA set, the maximal number of copies for a tRNA gene was randomly drawn from a uniformed distribution in the interval [1 30] since in the species analyzed, the maximal copy number for a tRNA gene is around 30. After choosing the maximal number of copies, a copy number was drawn for each tRNA gene from the interval [1 max]. To create a coding sequence, the maximal length of a gene, in amino acids, was drawn from a uniformed distribution in the interval [200 400]. A random codon was then chosen for each position in the sequence. In addition, for the purpose of the tAI normalization, a random "genome codon usage" was generated from a uniformed

distribution. The tAI for each of the two genes in the set and the tAI decomposition was then calculated.

4.1.6. Gene Ontology (GO) data

(Section 2.1.3 in Results) A file containing *S. cerevisiae* genes and their GO association was supplied by Ophir Shalem from my lab. In brief, The Gene Ontology (GO) database (Harris et al. 2004) was downloaded from http://www.geneontology.org. A file relating each *S. cerevisiae* ORF to GO terms (gene association file) was downloaded from SGD (Balakrishnan et al.). For the purpose of avoiding redundant statistical tests, a non-redundant group of GO terms was created by removing categories with a correlation ≥ 0.9 between their genes and that of another category. Each GO term was considered to annotate any orthologous group containing a *S. cerevisiae* gene that is associated with this term.

4.1.7. Statistical analyses

4.1.7.1. Cluster analysis

(Sections 2.1.3 in Results) K-means clustering of the tAI decomposition for each organism pair was performed using the MATLAB/Math Works Inc. statistical package. Euclidean Distance was used as a distance measure. To prevent biases which might result from a wrong choice of starting points, each clustering process was repeated using 500 different choices of starting points, and the output that minimized the sum of distances of a point to its cluster center was chosen. The number of clusters was chosen based on (Ray and Turi 1999). Briefly, the optimal number of clusters is chosen by minimizing the ratio between the intra cluster distance, which is the distance of a point to its cluster center, and the inter cluster distance, which is the distance between clusters centers.

The option to use hierarchical clustering was also examined, however, the clustering outputs were usually very unbalanced, with a few very large clusters and many small clusters, and the clusters seemed non-homogeneous, i.e. each cluster contained several profiles of decompositions.

4.1.7.2. Calculation of functional enrichment for clusters

(Sections 2.1.3 in Results) In each cluster the enrichment of each of the non-redundant GO-terms was checked. Enrichment was assessed using the one-sided hypergeometric test, and was corrected for multiple testing using the False Discovery

Rate (FDR) method (Benjamini and Hochberg 1995) with an FDR of 5%. Since GO annotation were assigned based on *S. cerevisiae* genes, for the enrichment analysis, only genes with an ortholog in *S. cerevisiae* were used.

4.2. Methods used for The effect of viral tRNA genes on the translation efficiency of viruses and their hosts

4.2.1. Protein and Coding sequences

For all viral and host species in the analysis, protein and coding sequences were downloaded from GenBank (http://www.ncbi.nlm.nih.gov/Genbank). The table below contains a list of all species, and their corresponding RefSeq accession number:

Species Name	**RefSeq accession number**
Synechococcus sp. WH8102	NC_005070
Synechococcus sp. WH7803	NC_009481
Prochlorococcus marinus str. MIT 9313	NC_005071
Prochlorococcus marinus str. MIT 9303	NC_008820
Prochlorococcus marinus str. MIT 9211	NC_009976
Prochlorococcus marinus str. MIT 9215	NC_009840
Prochlorococcus marinus str. MIT 9312	NC_007577
Prochlorococcus marinus str. MIT 9515	NC_008817
Prochlorococcus marinus str. NATL1A	NC_008819
Prochlorococcus marinus str. NATL2A	NC_007335
Prochlorococcus marinus str. SS120	NC_005042
Prochlorococcus marinus str. MED4	NC_005072
Synechococcus phage syn9	NC_008296

4.2.2. tRNA gene copy numbers in viruses and their hosts

tRNA gene copy numbers were obtained by applying the tRNAscan-SE software version 1.23 (Lowe and Eddy 1997). Chromosomal sequences for the species analyzed were downloaded from GenBank (http://www.ncbi.nlm.nih.gov/Genbank).

4.2.3. Calculation of the tRNA adaptation index (tAI) for coding sequences

Calculation of the tAI for the coding sequences was performed as explained in section 4.1.3. For each organism, the calculation was done once by including the viral genes in the host genome without using the viral tRNA genes as part of the cellular tRNA pool, and once by including the viral genes and using the viral tRNA genes as part of the cellular tRNA pool. The tAI difference for every gene was then calculated by taking the log of the ratio between the tAI which incorporates the viral tRNA and the tAI which does not. Note that due to the normalization, the tAI of a gene can either be improved or decreased. See discussion for details.

4.2.4. Assignment of genes to functional groups

(Section 2.2.1 in Results) Each of the virus Syn9 genes was manually assigned (by Keren Limor Waisberg from the Schertz lab, unpublished data) to one of six functional groups based on their annotation in (Peter R. Weigele 2007).

(Section 2.2.2 in Results) Assignment of the WH8102 genes to functional groups was downloaded from the CMR website (Peterson et al. 2001). Genes with no predicted function were assigned to one of four categories. "No Data" assigned proteins are hypothetical proteins with no homolog in other organisms (defined as "hypothetical proteins" in the CMR annotation). "Hypothetical Proteins" are proteins with homologs in other organisms, which are also hypothetical (defined as "Conserved Hypothetical" in the CMR annotation)."Unknown Function" assigned proteins are proteins with significant similarity to genes in other organisms, but which their function is unknown, while "Unclassified" assigned proteins are protein with no assigned function.

4.2.5. Statistical analysis

4.2.5.1. Clustering of the viral genes tAI difference across species

(Section 2.2.1 in Results) A matrix containing the tAI ratio of each viral gene (columns) across all the bacteria (rows) was generated and clustered using hierarchical clustering along the rows and then along the columns. Both clustering processes were used with Euclidean Distance as the distance measure and averaged linkage as the linkage measure.

4.2.5.2. Analysis of differences in the tAI ratio

For all the analyses that tested for difference in the tAI ratio across groups [differences between viral genes functional groups, differences between hosts and differences between host functional groups, (Section 2.2.1, 2.2.2 in Results)], the Kruskal-Wallis test (Rice 1995), a non-parametric analog of the one-way analysis of variance (ANOVA) was used in order to test for a difference of the median tAI ratio between groups. The results were further tested to find the source of difference in medians, using Wilcoxon rank-sum test (Rice 1995) and corrected for multiple tests using an FDR (Benjamini and Hochberg 1995) of 5%.

4.2.5.3. Calculation of functional enrichment for highly elevated genes

(Section 2.2.2 in Results) The 200 genes with the most elevated tAI upon introducing the viral tRNA genes were chosen for enrichment analysis. The enrichments were assessed using one-sided hypergeometric test, corrected for multiple testing using an FDR (Benjamini and Hochberg 1995) of 20%.

4.2.6. Generating the random tRNA sets

(Section 2.2.3 in Results) The tRNA sets were created by randomly choosing six anti-codons, allowing repetitions in each set. This process was repeated 10,000 times. Only non-repetitive sets were retained.

4.3. Methods used for analyses of the spatial patterns in the translation efficiency

4.3.1. Coding sequences and tRNA gene copy numbers

Coding sequences and the tRNA gene copy numbers used were the same as described in section 4.1.1 and 4.1.2, with the addition of *E.coli*, which was obtained as described in 4.1.1 and 4.1.2 for the yeast species.

4.3.2. Calculation of the local tAI profile

In order to create the local profile, the weight per codon, namely the tAI value for each individual codon, was calculated and normalized as described in (dos Reis et al. 2004). Each gene was then assigned a tAI profile which was simply the sequence of its codons' weights, omitting the start codon.

The averaged genome profile was calculated by lining up all the genes once according to the start codon, and once according to the stop codon, and averaging the weights along positions of the sequences.

For the control, each sequence was shuffled, and the average genome profile was calculated. This process was repeated 100 times. The mean and standard deviation of the 100 sets of profile was then calculated for each position.

For the Expected tAI (EtAI) profile, all codons which belong to the same amino acid were assigned the same weight, which is the weighted average of the original codons' tAI, weighted by the relative codon usage, i.e.:

$$\mathrm{EtAI(AA)} = \sum_{Ci \in codons(AA)} codon\_usage(Ci) \times tAI(Ci)$$

$$codon\_usage(Ci) = \#Ci \Big/ \sum_{Cj \in codons(AA)} \#Cj$$

The averaged genome profile was then calculated as described above.

For the Averaged tAI profile, all codons which belong to the same amino acid were assigned the same weight, which is the normal arithmetic average of the original AA codons' normalized tAIs, i.e.:

$$tAI(AA) = \sum_{Ci \in codons(AA)} tAI(Ci) \Big/ \left| codons(AA) \right| \ .$$

4.3.3. Generating the GC randomized sequences

(Section 2.3.1 figure 2 in Results) In order to randomize each sequence, but retain the local GC content at each codon, the codons were divided into 10 groups, according to the number of times G or C appear in them. For each sequence, every codon was replaced by a randomly chosen codon from the same group of the original codon. The local tAI profile and the averaged genome profile were then calculated as above.

4.3.4. Simulation parameters

The simulation was run 1,000 times on a 150 codons long mRNA, comprised of 30 "slow" codons, 30 "fast" codons and 90 "medium" codons, each run with a different layout of the codons which was randomly permuted.

The simulation time step was 0.001 sec, and each run included 1,800,000 steps, which correspond to 30 minutes.

The rates used were as follows:

Wb – 1 sec^{-1}

Wm-fast – 10 sec^{-1}

Wm-medium – 2 sec^{-1}

Wm-slow – 1 sec^{-1}.

Runs on real sequences were performed with the same rate parameters, on sequences that are at least 200 codons in length. Each codon was assigned to a rate category ("Slow", "Medium" or "Fast") based on its tAI. Bottom and top 25% codons were assigned a "Slow" and "Fast" rate respectively. The rest of the codons were assigned a "Medium" rate.

5. Literature Cited

Adam, E.-W. (1996). "The Close Proximity of Escherichia coli Genes: Consequences for Stop Codon and Synonymous Codon Use." Journal of Molecular Evolution **42**(2): 73-78.

Akashi, H. (2001). "Gene expression and molecular evolution." Curr Opin Genet Dev **11**(6): 660-6.

Akashi, H. (2003). "Translational Selection and Yeast Proteome Evolution." Genetics **164**(4): 1291-1303.

Alexeyenko, A., I. Tamas, et al. (2006). "Automatic clustering of orthologs and inparalogs shared by multiple proteomes." Bioinformatics **22**(14): e9-15.

Arava, Y., Y. Wang, et al. (2003). "Genome-wide analysis of mRNA translation profiles in Saccharomyces cerevisiae." Proc Natl Acad Sci U S A **100**(7): 3889-94.

Arnaud, M. B., M. C. Costanzo, et al. (2005). "The Candida Genome Database (CGD), a community resource for Candida albicans gene and protein information." Nucleic Acids Res **33**(Database issue): D358-63.

Bailly-Bechet, M., M. Vergassola, et al. (2007). "Causes for the intriguing presence of tRNAs in phages." Genome Research **17**(10): 1486-1495.

Balakrishnan, R., K. R. Christie, et al. "Saccharomyces Genome Database" ftp://ftp.yeastgenome.org/yeast/

Barnett, J. A. and K. D. Entian (2005). "A history of research on yeasts 9: regulation of sugar metabolism." Yeast **22**(11): 835-94.

Benjamini, Y. and Y. Hochberg (1995). "Controlling the False Discovery Rate - a Practical and Powerful Approach to Multiple Testing." Journal of the Royal Statistical Society Series B-Methodological **57**(1): 289-300.

Berbee, M. and J. Taylor (2001). Systematics and evolution. The Mycota. D. McLaughlin, E. McLaughlin and P. Lemke. Berlin, Springer. **VIIB:** 229-245.

Blanga-Kanfi, S., M. Amitsur, et al. (2006). "PrrC-anticodon nuclease: functional organization of a prototypical bacterial restriction RNase." Nucl. Acids Res. **34**(11): 3209-3219.

Boutros, P. C. and A. B. Okey (2005). "Unsupervised pattern recognition: An introduction to the whys and wherefores of clustering microarray data." Brief Bioinform **6**(4): 331-343.

Braun, B. R., M. van Het Hoog, et al. (2005). "A human-curated annotation of the Candida albicans genome." PLoS Genet **1**(1): 36-57.

Broida, J. and J. Abelson (1985). "Sequence organization and control of transcription in the bacteriophage T4 tRNA region." Journal of Molecular Biology **185**(3): 545-563.

Bulmer, M. (1988). "Codon usage and intragenic position." Journal of Theoretical Biology **133**(1): 67-71.

Canchaya, C., G. Fournous, et al. (2004). "The impact of prophages on bacterial chromosomes." Molecular Microbiology **53**(1): 9-18.

Chen, G.-F. T. and M. Inouye (1990). "Suppression of the negative effect of minor arginine codons on gene expression; preferential usage of minor codons within

the first 25 codons of the Escherichia coli genes." Nucl. Acids Res. **18**(6): 1465-1473.

Chen, G. T. and M. Inouye (1994). "Role of the AGA/AGG codons, the rarest codons in global gene expression in Escherichia coli." Genes Dev. **8**(21): 2641-2652.

Clokie, M. R. J., J. Shan, et al. (2006). "Transcription of a 'photosynthetic' T4-type phage during infection of a marine cyanobacterium." Environmental Microbiology **8**(5): 827-835.

Daubin, V. and H. Ochman (2004). "Bacterial Genomes as New Gene Homes: The Genealogy of ORFans in E. coli." Genome Research **14**(6): 1036-1042.

dos Reis, M., R. Savva, et al. (2004). "Solving the riddle of codon usage preferences: a test for translational selection." Nucleic Acids Res **32**(17): 5036-44.

Drummond, D. A. and C. O. Wilke (2008). "Mistranslation-Induced Protein Misfolding as a Dominant Constraint on Coding-Sequence Evolution." **134**(2): 341-352.

Dujon, B., D. Sherman, et al. (2004). "Genome evolution in yeasts." Nature **430**(6995): 35-44.

Eyre-Walker, A. and M. Bulmer (1993). "Reduced synonymous substitution rate at the start of enterobacterial genes." Nucl. Acids Res. **21**(19): 4599-4603.

Galagan, J. E., S. E. Calvo, et al. (2005). "Sequencing of Aspergillus nidulans and comparative analysis with A. fumigatus and A. oryzae." Nature **438**(7071): 1105-15.

Gilad, Y., A. Oshlack, et al. (2006). "Expression profiling in primates reveals a rapid evolution of human transcription factors." Nature **440**(7081): 242-5.

Gray, N. K. and M. W. Hentze (1994). "Regulation of protein synthesis by mRNA structure." Molecular Biology Reports **19**(3): 195-200.

Groisman, E. A. and H. Ochman (1996). "Pathogenicity islands: Bacterial evolution in quantum leaps." Cell **87**(5): 791-794.

Harris, M. A., J. Clark, et al. (2004). "The Gene Ontology (GO) database and informatics resource." Nucleic Acids Res **32**(Database issue): D258-61.

Hertz-Fowler, C., C. S. Peacock, et al. (2004). "GeneDB: a resource for prokaryotic and eukaryotic organisms." Nucleic Acids Res **32**(Database issue): D339-43.

Ihmels, J., S. Bergmann, et al. (2005). "Rewiring of the Yeast Transcriptional Network Through the Evolution of Motif Usage." Science **309**(5736): 938-940.

Ikemura, T. (1982). "Correlation between the abundance of yeast transfer RNAs and the occurrence of the respective codons in protein genes. Differences in synonymous codon choice patterns of yeast and Escherichia coli with reference to the abundance of isoaccepting transfer RNAs." J Mol Biol **158**(4): 573-97.

Kanaya, S., Y. Yamada, et al. (1999). "Studies of codon usage and tRNA genes of 18 unicellular organisms and quantification of Bacillus subtilis tRNAs: gene expression level and species-specific diversity of codon usage based on multivariate analysis." Gene **238**(1): 143-55.

Kanz, C., P. Aldebert, et al. (2005). "The EMBL Nucleotide Sequence Database." Nucleic Acids Res **33**(Database issue): D29-33.

Kapp, L. D. and J. R. Lorsch (2004). "THE MOLECULAR MECHANICS OF EUKARYOTIC TRANSLATION." Annual Review of Biochemistry **73**(1): 657-704.

Kaufmann, G. (2000). "Anticodon nucleases." Trends in Biochemical Sciences **25**(2): 70-74.

Kimchi-Sarfaty, C., J. M. Oh, et al. (2007). "A "Silent" Polymorphism in the MDR1 Gene Changes Substrate Specificity." Science **315**(5811): 525-528.

Kliman, R. M. and A. Eyre-Walker (1998). "Patterns of Base Composition Within the Genes of Drosophila melanogaster." Journal of Molecular Evolution **46**(5): 534-541.

Kunisawa, T. (1992). "Synonymous codon preferences in bacteriophage T4: A distinctive use of transfer RNAs from T4 and from its host Escherichia coli." Journal of Theoretical Biology **159**(3): 287-298.

Kunisawa, T. (2000). "Functional Role of Mycobacteriophage Transfer RNAs." Journal of Theoretical Biology **205**(1): 167-170.

Kurland, C. G. (1992). "Translational accuracy and the fitness of bacteria." Annu Rev Genet **26**: 29-50.

Lawrence, J. G. and H. Ochman (1998). "Molecular archaeology of the Escherichia coli genome." Proceedings of the National Academy of Sciences of the United States of America **95**(16): 9413-9417.

Liljenstrom, H. and G. Vonheijne (1987). "Translation Rate Modification by Preferential Codon Usage - Intragenic Position Effects." Journal of Theoretical Biology **124**(1): 43-55.

Lowe, T. M. and S. R. Eddy (1997). "tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence." Nucleic Acids Res **25**(5): 955-64.

Lu, J. and C. Deutsch (2008). "Electrostatics in the Ribosomal Tunnel Modulate Chain Elongation Rates." Journal of Molecular Biology **384**(1): 73-86.

Lu, M.-J. and U. Henning (1994). "Superinfection exclusion by T-even-type coliphages." Trends in Microbiology **2**(4): 137-139.

Man, O. and Y. Pilpel (2007). "Differential translation efficiency of orthologous genes is involved in phenotypic divergence of yeast species." Nat Genet **39**(3): 415-21.

Mann, N. H., M. R. J. Clokie, et al. (2005). "The Genome of S-PM2, a "Photosynthetic" T4-Type Bacteriophage That Infects Marine Synechococcus Strains." J. Bacteriol. **187**(9): 3188-3200.

Mesyanzhinov, V. V., M. Karl, et al. (2004). Bacteriophage T4: Structure, Assembly, and Initiation Infection Studied in Three Dimensions. Advances in Virus Research, Academic Press. **Volume 63:** 287-352.

Mitarai, N., K. Sneppen, et al. (2008). "Ribosome collisions and translation efficiency: optimization by codon usage and mRNA destabilization." J Mol Biol **382**(1): 236-45.

Monger, B. C., M. R. Landry, et al. (1999). "Feeding selection of heterotrophic marine nanoflagellates based on the surface hydrophobicity of their picoplankton prey." Limnology and Oceanography **44**(8): 1917-1927.

Nishida, K., T. Kawasaki, et al. (1999). "Aminoacylation of tRNAs Encoded by Chlorella Virus CVK2." Virology **263**(1): 220-229.

Novembre, J. A. (2002). "Accounting for background nucleotide composition when measuring codon usage bias." Mol Biol Evol **19**(8): 1390-4.

Ohno, H., H. Sakai, et al. (2001). "Preferential usage of some minor codons in bacteria." Gene **276**(1-2): 107-115.

Palenik, B., B. Brahamsha, et al. (2003). "The genome of a motile marine Synechococcus." Nature **424**(6952): 1037-1042.

Percudani, R., A. Pavesi, et al. (1997). "Transfer RNA gene redundancy and translational selection in Saccharomyces cerevisiae." J Mol Biol 268(2): 322-30.

Powers, D. A. and P. M. Schulte (1998). "Evolutionary adaptations of gene structure and expression in natural populations in relation to a changing environment: A multidisciplinary approach to address the million-year saga of a small fish." The Journal of Experimental Zoology **282**(1-2): 71-94.

Peterson, J. D., L. A. Umayam, et al. (2001). "The Comprehensive Microbial Resource." Nucl. Acids Res. **29**(1): 123-125.

Prud'homme, B., N. Gompel, et al. (2007). "Emerging principles of regulatory evolution." Proceedings of the National Academy of Sciences **104**(Suppl 1): 8605-8612.

Pruess, M., P. Kersey, et al. (2005). "The Integr8 project--a resource for genomic and proteomic data." In Silico Biol **5**(2): 179-85.

Qin, H., W. B. Wu, et al. (2004). "Intragenic spatial patterns of codon usage bias in prokaryotic and eukaryotic genomes." Genetics **168**(4): 2245-60.

Ray, S. and R. H. Turi (1999). Determination of number of clusters in k-means clustering and application in colour image segmentation. 4th International Conference on Advances in Pattern Recognition and Digital Techniques, Calcutta, India, Narosa Publishing House.

Remm, M., C. E. Storm, et al. (2001). "Automatic clustering of orthologs and in-paralogs from pairwise species comparisons." J Mol Biol **314**(5): 1041-52.

Rice, J. (1995). Mathematical statistics and data analysis. Belmont, California, Duxbury Press (Wadsworth Publishing Company).

Rocap, G., D. L. Distel, et al. (2002). "Resolution of Prochlorococcus and Synechococcus Ecotypes by Using 16S-23S Ribosomal DNA Internal Transcribed Spacer Sequences." Appl. Environ. Microbiol. **68**(3): 1180-1191.

Rocha, E. P. C. and A. Danchin (2002). "Base composition bias might result from competition for metabolic resources." Trends in Genetics **18**(6): 291-294.

Segal, E., Y. Fondufe-Mittendorf, et al. (2006). "A genomic code for nucleosome positioning." Nature.

Sharp, P. M. and W. H. Li (1986). "An evolutionary perspective on synonymous codon usage in unicellular organisms." J Mol Evol **24**(1-2): 28-38.

Sharp, P. M. and W. H. Li (1987). "The codon Adaptation Index--a measure of directional synonymous codon usage bias, and its potential applications." Nucleic Acids Res **15**(3): 1281-95.

Souciet, J.-L., M. Aigle, et al. (2000). "Genomic Exploration of the Hemiascomycetous Yeasts: 1. A set of yeast species for molecular evolution studies." FEBS Letters **487**(1): 3-12.

Sullivan, M. B., J. B. Waterbury, et al. (2003). "Cyanophages infecting the oceanic cyanobacterium Prochlorococcus." Nature **424**(6952): 1047-1051.

Tan, Y., K. Zhang, et al. (2007). "Whole genome sequencing of a novel temperate bacteriophage of P.aeruginosa: evidence of tRNA gene mediating integration

of the phage genome into the host bacterial chromosome." Cellular Microbiology **9**(2): 479-491.

Varenne, S., J. Buc, et al. (1984). "Translation is a non-uniform process : Effect of tRNA availability on the rate of elongation of nascent polypeptide chains." Journal of Molecular Biology **180**(3): 549-576.

Vervoort, E. B., A. v. Ravestein, et al. (2000). "Optimizing heterologous expression in Dictyostelium: importance of 5' codon adaptation." Nucl. Acids Res. **28**(10): 2069-2074.

Wang, D., H.-M. Sung, et al. (2007). "Expression evolution in yeast genes of single-input modules is mainly due to changes in trans-acting factors." Genome Research **17**(8): 1161-1169.

Waterbury, J. B. and F. W. Valois (1993). "Resistance to Co-Occurring Phages Enables Marine Synechococcus Communities To Coexist with Cyanophages Abundant in Seawater." Appl. Environ. Microbiol. **59**(10): 3393-3399.

Weigele, P. R., W. H. Pope, et al. (2007). "Genomic and structural analysis of Syn9, a cyanophage infecting marine Prochlorococcus and Synechococcus." Environmental Microbiology **9**(7): 1675-1695.

Weiss, S. B., W. T. Hsu, et al. (1968). "Transfer RNA coded by the T4 bacteriophage genome." Proceedings of the National Academy of Sciences of the United States of America **61**(1): 114-121.

Widmann, M., M. Clairo, et al. (2008). "Analysis of the distribution of functionally relevant rare codons." BMC Genomics **9**: 207.

Wilson, J. H. (1973). "Function of the bacteriophage T4 transfer RNA's." Journal of Molecular Biology **74**(4): 753-754.

Wittkopp, P. J., B. K. Haerum, et al. (2004). "Evolutionary changes in cis and trans gene regulation." Nature **430**(6995): 85-88.

Wright, F. (1990). "The 'effective number of codons' used in a gene." Gene **87**(1): 23-9.

Zhang, S., E. Goldman, et al. (1994). "Clustering of low usage codons and ribosome movement." J Theor Biol **170**(4): 339-54.

6. Appendices

6.1. Appendix 1 – Distribution of the tAI components in the 28 yeast pairs tested

	trans		cis		co-evolution		tAI change	
pair	mean	std	mean	std	mean	std	mean	std
AN/CA	0.0004	0.0636	-0.0177	-0.0172	0.123	0.122	0.0144	0.0135
AN/CG	-0.001	0.0484	-0.1026	-0.1027	0.1068	0.1068	0.1024	0.1024
AN/DH	-0.0059	0.0652	-0.0758	-0.0799	0.1228	0.1201	0.0596	0.0608
AN/KL	-0.0044	0.0433	-0.0787	-0.0789	0.0972	0.0967	0.0837	0.084
AN/SP	-0.0338	0.0158	0.0615	0.0616	0.079	0.0778	-0.0099	-0.0102
AN/YL	0.076	0.0633	-0.2395	-0.2438	0.1142	0.1115	0.1412	0.1418
CA/CG	-0.0096	0.0506	0.0828	0.0829	0.1255	0.1255	-0.0727	-0.0728
CA/DH	-0.014	0.0189	0.0126	0.0084	0.1069	0.1019	-0.0148	-0.0152
CA/KL	-0.0069	0.0474	0.0801	0.0807	0.1201	0.1194	-0.0721	-0.0726
CA/SP	0.0175	0.0569	0.0612	0.0621	0.1189	0.1185	-0.0614	-0.0611
CA/YL	0.0173	0.066	-0.1606	-0.1588	0.1787	0.1803	0.12	0.118
CG/DH	0.0096	0.0483	-0.1021	-0.1022	0.128	0.1279	0.0711	0.0712
CG/KL	0.0019	0.015	-0.006	-0.0059	0.0846	0.0847	0.0044	0.0044
CG/SP	0.0013	0.0339	0.1006	0.1007	0.099	0.0991	-0.0838	-0.0837
CG/YL	0.0415	0.0644	-0.0109	-0.0107	0.1834	0.1838	-0.0565	-0.0566
DH/KL	-0.005	0.0477	0.087	0.0889	0.125	0.1236	-0.0612	-0.0621
DH/SP	0.0261	0.0565	0.0956	0.0986	0.1255	0.1244	-0.0821	-0.0827
DH/YL	0.053	0.0586	-0.0733	-0.0692	0.1715	0.1713	0.0167	0.0141
KL/SP	0.0017	0.0368	0.0911	0.0917	0.0947	0.0944	-0.0759	-0.0759
KL/YL	0.0491	0.0634	-0.044	-0.0434	0.1732	0.1737	-0.0312	-0.0317
SC/AN	0.004	0.0513	0.083	0.083	0.1007	0.1007	-0.0882	-0.0882
SC/CA	0.0332	0.0435	-0.1088	-0.1088	0.1295	0.1295	0.0748	0.0748
SC/CG	0.0141	0.018	-0.0156	-0.0156	0.0853	0.0853	0.0017	0.0017
SC/DH	0.0259	0.0437	-0.1083	-0.1083	0.1303	0.1303	0.0602	0.0602
SC/KL	0.0206	0.0147	-0.0218	-0.0218	0.0854	0.0854	0.0018	0.0018
SC/SP	0.0258	0.0409	0.0748	0.0748	0.091	0.091	-0.0824	-0.0824
SC/YL	0.046	0.0696	-0.0452	-0.0452	0.1696	0.1696	-0.0283	-0.0283
SP/YL	0.095	0.0595	-0.3198	-0.3246	0.1293	0.1272	0.1872	0.1884
random data	0.0001	0.0666	0	0.0627	0	0.0373	0.0001	0.0951

6.2. Appendix 2 – Results of enrichment tests for the glucose repression phenotype related categories

In each cluster the enrichment of each of the non-redundant GO-terms was checked. Enrichment was assessed using the one-sided hypergeometric test, and was corrected for multiple testing using the False Discovery Rate (FDR) method (Benjamini and Hochberg 1995) with an FDR of 5%. I extracted the clusters which are enriched for the related categories from all pairs where one yeast species shows the phenotype and the other does not.

GO category	Pair	Genes in category	Genes in cluster	Cluster size	P value
"Cytosolic part"	SC/AN	109	9	48	$3.59*10^{-05}$
			11	70	$2.72*10^{-05}$
			35	95	$1.64*10^{-28}$
	SC/CA	110	22	181	$7.55*10^{-09}$
	SC/DH	106	36	160	$5.09*10^{-13}$
	SC/KL	108	9	78	$1.53*10^{-04}$
			29	120	$3.58*10^{-12}$
	SC/YL	97	20	39	$6.02*10^{-12}$
			8	10	$2.37*10^{-11}$
			13	75	$2.04*10^{-07}$
			23	34	$3.88*10^{-12}$
	CG/AN	102	26	60	$3.72*10^{-13}$
			22	104	$3.21*10^{-12}$
	CG/CA	104	23	154	$3.63*10^{-11}$
			16	161	$1.23*10^{-05}$
	CG/DH	99	21	162	$1.91*10^{-09}$
	CG/KL	103	35	77	$1.41*10^{-12}$
			20	218	$4.16*10^{-07}$
	CG/YL	93	23	37	$2.00*10^{-27}$
			8	10	$1.96*10^{-11}$
			12	95	$1.72*10^{-05}$
			19	45	$1.96*10^{-18}$
	SP/CA	95	36	195	$7.87*10^{-13}$
	SP/DH	92	28	148	$8.34*10^{-15}$
			24	50	$2.08*10^{-23}$
	SP/KL	94	24	192	$6.89*10^{-09}$
			16	192	$5.29*10^{-04}$
	SP/YL	84	16	91	$1.64*10^{-08}$
"Organellar ribosome"	SC/AN	55	9	77	$6.51*10^{-06}$
	SC/CA	74	9	103	$1.74*10^{-04}$
	SC/DH	69	16	173	$7.01*10^{-08}$

			15	74	$2.13*10^{-12}$
	SC/KL	79	13	104	$4.51*10^{-08}$
			7	81	$7.51*10^{-04}$
	SC/YL	67	11	117	$2.10*10^{-05}$
			17	104	$1.00*10^{-11}$
			7	36	$6.38*10^{-06}$
	CG/AN	53	11	55	$1.53*10^{-09}$
	CG/DH	66	10	105	$2.01*10^{-05}$
			11	85	$3.29*10^{-07}$
	CG/KL	76	6	55	$5.36*10^{-04}$
			8	76	$7.84*10^{-05}$
	CG/YL	67	8	39	$1.07*10^{-06}$
			18	118	$2.05*10^{-11}$
			10	81	$5.92*10^{-06}$
	SP/AN	45	6	46	$1.26*10^{-04}$
			8	60	$6.90*10^{-06}$
	SP/CA	54	7	24	$2.21*10^{-07}$
	SP/DH	50	12	126	$1.94*10^{-06}$
			7	15	$3.05*10^{-09}$
	SP/KL	54	8	51	$3.94*10^{-06}$
			7	47	$2.40*10^{-05}$
	SP/YL	49	10	93	$6.49*10^{-06}$
			11	102	$2.00*10^{-06}$
			14	46	$1.90*10^{-14}$
"Glycolysis"	SC/AN	11	7	70	$7.6*10^{-10}$
	SC/CA	14	6	181	$2.74*10^{-05}$
	SC/DH	13	8	160	$1.24*10^{-08}$
	SC/KL	15	7	78	$3.46*10^{-09}$
	SC/YL	11	4	75	$7.93*10^{-05}$
	CG/YL	11	4	95	0.000226
	SP/CA	10	5	195	0.00034
	SP/YL	9	4	49	$1.41*10^{-05}$
"Aerobic respiration"	SC/AN	52	8	105	0.000325
	SC/CA	61	8	103	0.000239
	SC/DH	54	8	74	$9.65*10^{-06}$
	SC/KL	67	7	55	$2.2*10^{-05}$
			12	81	$3.15*10^{-09}$
			7	36	$1.15*10^{-06}$
	SC/YL	54	10	104	$5.86*10^{-06}$
	CG/KL	67	9	55	$2.08*10^{-07}$
	CG/YL	54	5	39	0.000474
			10	118	$2.39*10^{-05}$
	SP/AN	47	7	46	$1.52*10^{-05}$
	SP/CA	49	8	40	$3.1*10^{-07}$
	SP/KL	50	12	47	$1.26*10^{-11}$

6.3. Appendix 3 – The tRNA repertoires of the bacteria species analyzed

The gene copy numbers of all tRNA species in 12 bacteria and one phage were determined using an HMM-based approach (Lowe and Eddy 1997). Rows that correspond to the seven tRNAs that are assumed to be absent in all living species (due to the structure of the genetic table and wobble interactions, which imply that their presence may result in mistranslation of some codons) are shown in red. The first three columns show, respectively, the anticodon borne by the tRNA, the codon that is perfectly decoded by the anticodon, and the amino acid that corresponds to the codon.

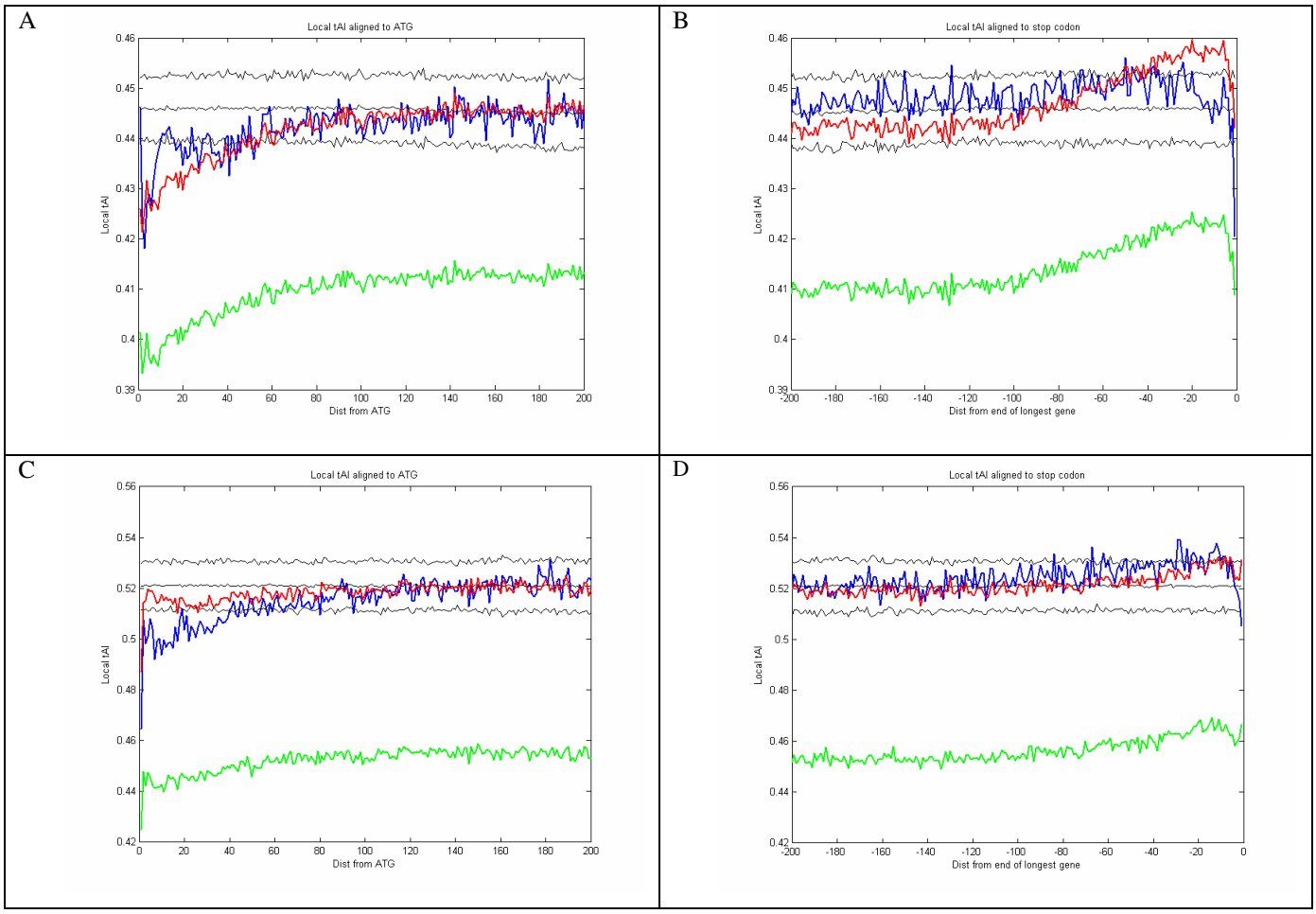
anticodon	codon	amino acid	med 4	mit 9211	mit 9215	mit 9303	mit 9312	mit 9313	mit 9515	natl 1a	natl 2a	ss 120	wh 7803	wh 8102	syn9
AAA	TTT	F	0	0	0	0	0	0	0	0	0	0	0	0	0
GAA	TTC	F	1	1	1	1	1	1	1	1	1	1	1	1	0
TAA	TTA	L	1	1	1	1	1	1	1	1	1	1	1	1	1
CAA	TTG	L	1	1	1	1	1	1	1	1	1	1	1	1	0
AGA	TCT	S	0	0	0	0	0	0	0	0	0	0	0	0	0
GGA	TCC	S	1	1	1	1	1	1	1	1	1	1	1	1	0
TGA	TCA	S	1	1	1	1	1	1	1	1	1	1	1	1	0
CGA	TCG	S	1	1	1	1	1	1	1	1	1	1	1	1	0
ATA	TAT	Y	0	0	0	0	0	0	0	0	0	0	0	0	0
GTA	TAC	Y	1	1	1	1	1	1	1	1	1	1	1	1	0
ACA	TGT	C	0	0	0	0	0	0	0	0	0	0	0	0	0
GCA	TGC	C	1	1	1	1	1	1	1	1	1	1	1	1	0
CCA	TGG	W	1	1	1	1	1	1	1	1	1	1	1	1	0
AAG	CTT	L	1	1	1	0	1	0	1	1	1	1	0	0	0
GAG	CTC	L	0	0	0	1	0	1	0	0	0	0	1	1	0
TAG	CTA	L	1	1	1	1	1	1	1	1	1	1	1	1	0
CAG	CTG	L	0	1	0	1	0	1	0	0	0	1	1	1	0
AGG	CCT	P	0	0	0	0	0	0	0	0	0	0	0	0	0
GGG	CCC	P	1	1	1	1	1	1	1	1	1	1	1	1	0
TGG	CCA	P	1	1	1	1	1	1	1	1	1	1	1	1	0
CGG	CCG	P	0	0	0	1	0	1	0	0	0	0	1	1	0
ATG	CAT	H	0	0	0	0	0	0	0	0	0	0	0	0	0
GTG	CAC	H	1	1	1	1	1	1	1	1	1	1	1	1	0
TTG	CAA	Q	1	1	1	1	1	1	1	1	1	1	1	1	0
CTG	CAG	Q	0	0	0	0	0	0	0	0	0	0	0	0	0
ACG	CGT	R	1	1	1	1	1	1	1	1	1	1	1	1	0
GCG	CGC	R	0	0	0	0	0	0	0	0	0	0	0	0	0
TCG	CGA	R	0	0	0	0	0	0	0	0	0	0	0	0	0
CCG	CGG	R	1	1	1	1	1	1	1	1	1	1	1	1	0
AAT	ATT	I	0	0	0	0	0	0	0	0	0	0	0	0	0
GAT	ATC	I	1	1	1	2	1	2	1	1	1	1	2	2	0
TAT	ATA	I	0	0	0	0	0	0	0	0	0	0	0	0	0
CAT	ATG	M	3	3	2	2	3	2	3	3	3	3	2	2	0
AGT	ACT	T	0	0	0	0	0	0	0	0	0	0	0	0	0
GGT	ACC	T	1	2	3	1	1	1	1	1	1	1	1	1	0
TGT	ACA	T	1	1	1	1	1	1	1	1	1	1	1	1	1
CGT	ACG	T	1	1	1	1	1	1	1	1	1	1	1	1	0
ATT	AAT	N	0	0	0	0	0	0	0	0	0	0	0	0	0
GTT	AAC	N	1	1	1	1	1	1	1	1	1	1	1	1	1
TTT	AAA	K	1	1	1	1	1	1	1	1	1	1	1	1	0

anticodon	codon	amino acid	med 4	mit 9211	mit 9215	mit 9303	mit 9312	mit 9313	mit 9515	natl 1a	natl 2a	ss 120	wh 7803	wh 8102	syn9
CTT	AAG	K	0	0	0	0	0	0	0	0	0	0	0	0	0
ACT	AGT	S	0	0	0	0	0	0	0	0	0	0	0	0	0
GCT	AGC	S	1	1	1	1	1	1	1	1	1	1	1	1	0
TCT	AGA	R	1	1	1	1	1	1	1	1	1	1	1	1	1
CCT	AGG	R	1	1	1	1	1	1	1	1	1	1	1	1	0
AAC	GTT	V	0	0	0	0	0	0	0	0	0	0	0	0	0
GAC	GTC	V	1	1	1	1	1	1	1	1	1	1	1	1	0
TAC	GTA	V	1	1	1	1	1	1	1	1	1	1	1	1	1
CAC	GTG	V	0	0	0	1	0	1	0	0	0	0	1	1	0
AGC	GCT	A	0	0	0	0	0	0	0	0	0	0	0	0	0
GGC	GCC	A	1	1	1	1	1	1	1	1	1	1	1	1	0
TGC	GCA	A	1	1	1	2	1	2	1	1	1	1	2	2	1
CGC	GCG	A	0	0	0	1	0	1	0	0	0	0	1	1	0
ATC	GAT	D	0	0	0	0	0	0	0	0	0	0	0	0	0
GTC	GAC	D	1	1	1	1	1	1	1	1	1	1	1	1	0
TTC	GAA	E	1	1	1	1	1	1	1	1	1	1	1	1	0
CTC	GAG	E	0	0	0	0	0	0	0	0	0	0	0	0	0
ACC	GGT	G	0	0	0	0	0	0	0	0	0	0	0	0	0
GCC	GGC	G	1	1	1	1	1	1	1	1	1	1	1	1	0
TCC	GGA	G	1	1	1	1	1	1	1	1	1	1	1	1	0
CCC	GGG	G	0	1	0	1	0	1	0	1	1	1	1	1	0
Total			37	40	38	43	37	43	37	38	38	39	43	43	6

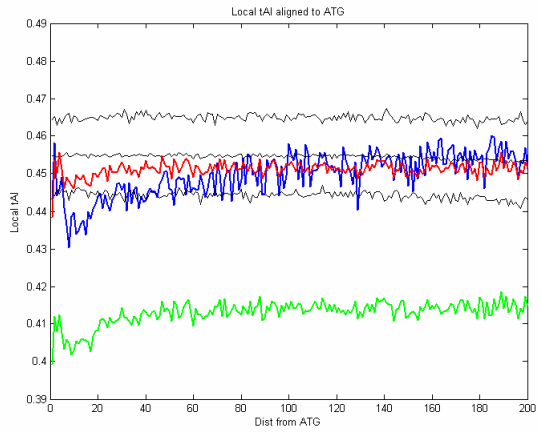
6.4. Appendix 4 – tAI profiles of six yeast species

The local tAI profile (blue), EtAI profile (red) and AA averaged tAI profile (green) including randomized profile +- 3 standard deviations are presented for six yeast species (see Methods).

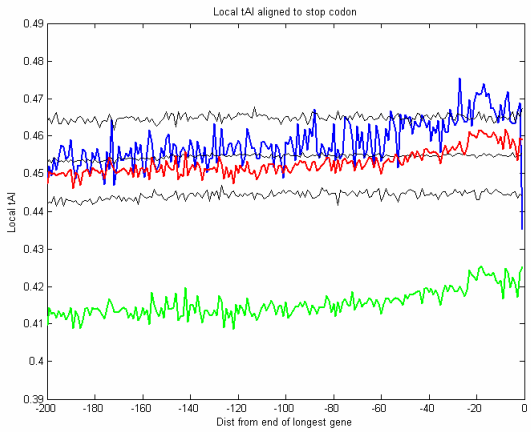
A,B – *A. nidulans* (start; end) C,D – *C. albicans* (start; end) E,F – *C. glabrata* (start; end) G,H – *D. hansenii* (start; end) I,J – *K. lactis* (start; end) K,L – *S. pombe* (start; end)



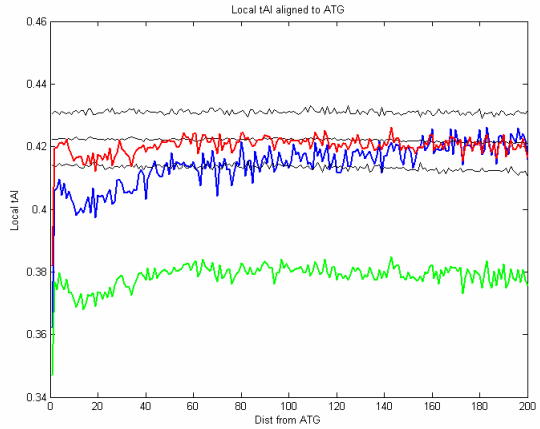
E


Local tAI aligned to ATG
Local tAI
Dist from ATG

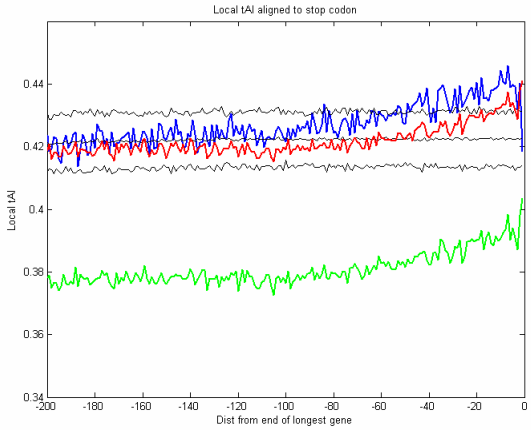
F


Local tAI aligned to stop codon
Local tAI
Dist from end of longest gene

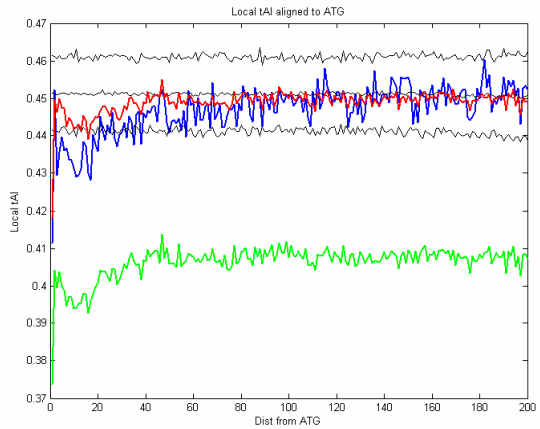
G


Local tAI aligned to ATG
Local tAI
Dist from ATG

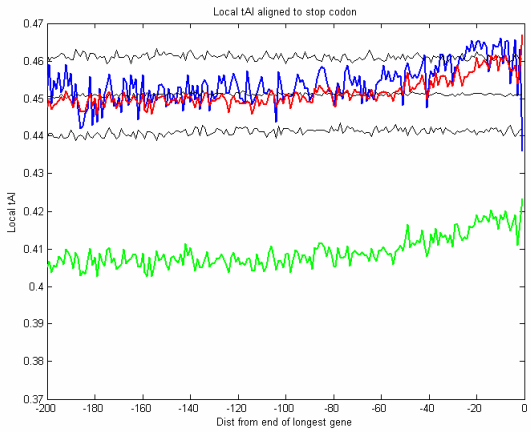
H


Local tAI aligned to stop codon
Local tAI
Dist from end of longest gene

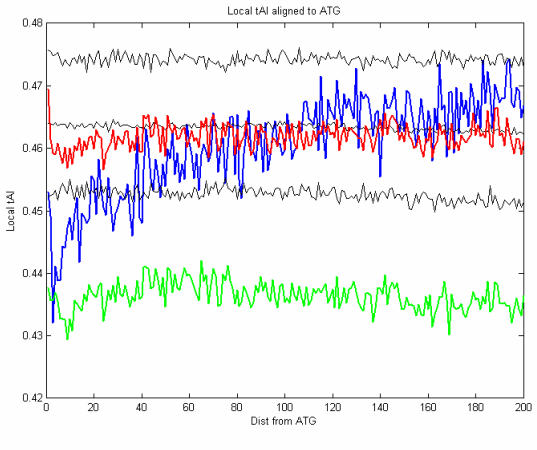
I


Local tAI aligned to ATG
Local tAI
Dist from ATG

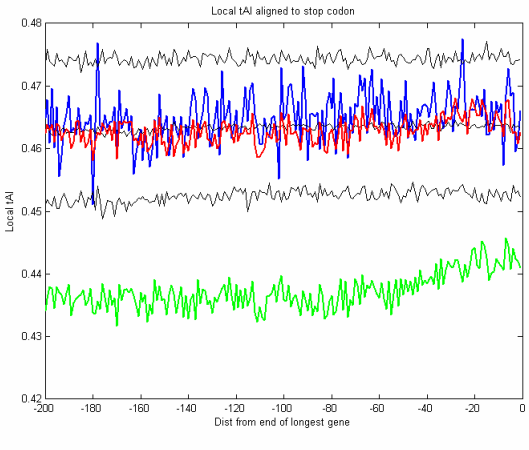
J


Local tAI aligned to stop codon
Local tAI
Dist from end of longest gene

K


Local tAI aligned to ATG
Local tAI
Dist from ATG
0.48
0.47
0.46
0.45
0.44
0.43
0.42
0
20
40
60
80
100
120
140
160
180
200

L


Local tAI aligned to stop codon
Local tAI
Dist from end of longest gene
0.48
0.47
0.46
0.45
0.44
0.43
0.42
-200
-180
-160
-140
-120
-100
-80
-60
-40
-20
0

6.5. Appendix 5 – description of the tAI measure

The tRNA Adaptation Index, tAI (dos Reis et al. 2004), uses the tRNA genes copy numbers (tGCNs) in the genome as a means to calculate the translation efficiency, by assigning weights to each codon based on abundance of its cognate tRNA taking into account wobble interactions.

The weight for each codon is calculated as follows:

$$W_i = \sum_{j=1}^{n_i} (1 - s_{ij}) \mathrm{tGCN}_{ij} \quad \textbf{(1)}$$

where n_i is the number of tRNA isoacceptors that recognize the *i*th codon, tGCN_{ij} is the gene copy number of the *j*th tRNA that recognize the *i*th codon, and s_{ij} is a selective constraint on the efficiency of the codon-anticodon interaction. The s_{ij} used in this study were taken from (dos Reis et al. 2004) and are 0 for all the perfect match interactions. The table below details the values for the wobble interactions:

Wobble interaction (anti-codon : codon)	S-value
G:U	0.41
I:C	0.28
I:A	0.9999
U:G	0.68
L:A (lysidine, for prokaryotes)	0.89

Each codon then is assigned a relative weight, w_i calculated based on W_i as:

$$w_i = W_i / W_{\max}, \; 0 < w_i \leq 1 \quad \textbf{(2)}$$

where W_{max} is the maximum W_i value.

The tAI_g of a gene is defined as the geometric mean of the relative weights of its codons:

$$\mathrm{tAI}_g = \left(\prod_{k=1}^{l_g} w_{i_{kg}} \right)^{1/l_g} \quad \textbf{(3)}$$

where i_{kg} is the codon defined by the *k*th triplet in gene g and l_g is the length of gene g without the stop codon.