Revisiting the CAI from a whole-genome perspective: analyzing the relationship between gene expression and codon occurrence in yeast using a variety of models

Ronald Jansen[1] & Mark Gerstein[1, 2, †]

Department of Molecular Biophysics & Biochemistry[1] and Computer Science[2]

266 Whitney Avenue, Yale University

PO Box 208114, New Haven, CT 06520

(203) 432-6105, FAX (360) 838-7861

ronald.jansen@yale.edu

mark.gerstein@yale.edu

Revised version submitted to *Nucleic Acids Research*, February 13, 2003

† Corresponding author

Abstract

Highly expressed genes in many bacteria and small eukaryotes often have a strong compositional bias, in terms of codon usage. Two widely used numerical indices, the codon adaptation index (CAI) and the codon usage, use this bias to predict the expression level of genes. Both indices are based on fairly simple assumptions about which genes are most highly expressed, which were known when they were first derived: the CAI was originally based on the codon composition of a set of only 24 highly expressed genes, and the codon usage, on assumptions about which functional classes of genes are highly expressed in fast-growing bacteria. Given the recent advent of genome-wide expression data, we should be able to improve on these assumptions. Here, we measure, in yeast, the degree to which consideration of the current genome-wide expression datasets improves the performance of both numerical indices. Indeed, we find that by changing the parameterization of each model its correlation with actual expression levels can be somewhat improved, although both indices are fairly insensitive to the exact way they are parameterized. This insensitivity indicates a consistent codon bias amongst highly expressed genes. We also attempt direct linear regression of codon composition against genome-wide expression levels (and protein abundance data). This has some similarity with the CAI formalism and yields an alternative model for the prediction of expression levels based on the coding sequences of genes. More information is at http://bioinfo.mbb.yale.edu/expression/codons.

Introduction

It is well known that highly expressed genes exhibit a strong bias for particular codons in many bacteria and small eukaryotes. One suggested explanation is the observation that there appears to be a relationship between tRNA abundance and codon bias (1-3). Several reviews on this topic have been published previously (4,5).

In 1987, the codon adaptation index (CAI) was proposed as a quantitative way of predicting the expression level of a gene based on its codon sequence (1). More recently, the "codon usage" was introduced as an alternative quantitative indicator (3). It also uses the occurrence of codons in a gene sequence to predict whether genes are likely to be highly expressed, although the formalism is quite different from the one used for the CAI. A related method, the codon bias formalism, is based on similar principles (6).

Expression level indicators such as these are widely used and are important in a variety of contexts. First, there is the annotation of genome sequences. The expression level indicators can serve as one of the variables to determine how likely the transcription and translation of an open reading frame (ORF) into a protein product is. Second, in heterologous gene expression, the codon-based expression indicators are helpful for finding the codon sequences that are most likely to yield high expression. The codon-based expression indicators and related methods are also often used as convenient "rules of thumb" in other applications.

Given that the codon-based expression models have these important applications, it is perhaps surprising that they are still based on rather qualitative assumptions about gene expression. For instance, the parameters underlying the CAI model rely on the codon composition of only a limited set of highly expressed genes; to define the parameters in the CAI model (see below), Sharp et al. counted the codon frequency in only 24 highly expressed genes. About half of these genes are ribosomal; the remaining ones are mostly metabolic enzymes (1).

In the codon usage model, the parameters are based on a somewhat broader set of highly expressed genes. The codon usage model has mainly been applied to fast growing

bacteria, for which, as Karlin et al. have shown, it is a reasonable assumption that ribosomal genes, chaperones, and translation processing factors are highly expressed (7,8).

In summary, the codon-based expression models are based on qualitative estimates of the expression levels of limited gene sets. But since these models were first proposed, several quantitative expression datasets, covering the majority of genes in a genome, have become available. This raises the natural question whether we could improve the parameters of the codon-based expression indicators by considering larger sets of genes with more accurate expression data. We present the results of such a procedure here, using the expression information available for the organism yeast.

In the following sections we briefly recap the CAI and codon usage formalisms. Later, we show how to calculate new parameters for these models. We also propose an alternative linear model to predict the expression levels from the codon composition of genes.

The CAI model

The CAI model assigns a parameter, termed "relative adaptiveness" by Sharp et al., to each of the 61 codons (stop codons excluded) (1). The relative adaptiveness of a codon is defined as its frequency relative to the most often used synonymous codon; note that this parameter is computed from a set of highly expressed genes *G* (we leave aside the question of how to define this set of genes for now). It is given by:

$$w_{aa,i}(G) = \frac{f_{aa,i}(G)}{f_{aa,\max}(G)} \qquad [1]$$

where $f_{aa,i}$ is the frequency of codon *i* (which encodes amino acid *aa*), and $f_{aa,max}$ the frequency of the codon most often used for encoding amino acid *aa* in a set of highly expressed genes *G*. The relative adaptiveness parameter $w_{aa,i}$ ranges from 0 to 1, with 0 indicating that a codon is not present at all in *G*, and 1, a codon that occurs most often in *G* for a given amino acid.

The CAI of a gene *g* is then simply the geometric average of the relative adaptiveness of all codons in a gene sequence:

$$CAI_g = \prod_{i=1}^{N_g} w_i^{1/N} \qquad [2]$$

Here, w_i is the relative adaptiveness of the *i*th codon in a gene with *N* codons. This formula can be transformed into:

$$CAI_g = \prod_{k=1}^{61} w_k^{X_{k,g}} \qquad [3]$$

where w_k now represents the relative adaptiveness of the *k*th out of the 61 codons in the genetic code (excluding stop codons); $X_{k,g}$ is the fraction of codon *k* among the total number of codons in gene *g*:

$$X_{k,g} = \frac{C_{k,g}}{\sum_{i=1}^{61} C_{i,g}} \qquad [4]$$

where $C_{k,g}$ is the number of times codon *k* appears in gene *g*. Note that $w_k = w_k(G)$ in equation [3] is dependent on the set of highly expressed genes *G*.

Like the relative adaptiveness, the CAI also ranges from 0 to 1. Higher CAI values indicate genes that are more likely to be highly expressed.

The codon usage model

Karlin et al. define the codon bias of a gene *g* relative to a gene set *G* as (4):

$$B(g \mid G) = \sum_{aa} p_{aa}(g) \left(\sum_{(x,y,z)=aa} \left| f(x,y,z) - g(x,y,z) \right| \right) \qquad [5]$$

where $p_{aa}(f)$ is the fraction of amino acid *aa* in gene *g*; $f(x, y, z)$ the frequency of a codon triplet (x, y, z) in gene *g* normalized such that $f(x, y, z) = 1$ if (x, y, z) is the most common

synonymous codon; $g(x, y, z)$ is the corresponding normalized codon frequency in gene set G. Equation [5] is written in the notation of Karlin et al. We can rewrite equation [5] in our own notation as follows:

$$B(g \mid G) = \sum_{k} \left| X_{k,g} - X_{k,G} \right| \qquad [6]$$

where $X_{k,g}$ and $X_{k,G}$ are defined as in equation [4]. Note that k has replaced (x, y, z) as the summation index. Given these definitions, Karlin et al. defined an expression level measure $E(g)$ as follows (8):

$$E(g) = \frac{B(g \mid C)}{\frac{1}{2} B(g \mid RP) + \frac{1}{4} B(g \mid Ch) + \frac{1}{4} B(g \mid Tf)} \qquad [7]$$

where the gene set C comprises all genes in the genome, RP the ribosomal proteins, Ch chaperones, and Tf translation processing factors. $E(g)$ is close to zero if gene g has a codon composition close to the average composition of the genome ($E(g)$ _ 0 because $B(g|C)$ _ 0), while $E(g)$ would take on very large values if the codon composition of gene g is close to the composition of ribosomal genes, chaperones and translation processing factors ($E(g) >> 1$ because $B(g|RP)$, $B(g|Ch)$, $B(g|Tf)$ _ 0). The idea is that highly expressed genes tend to have higher values of E than lowly expressed genes.

Karlin et al. have shown that highly expressed genes can best be differentiated from lowly expressed genes in the multidimensional space of the different codon bias terms $B(g|RP)$, $B(g|Ch)$ and $B(g|Tf)$ (8). However, in this study, we use the simplified expression measure $E(g|G)$, defined as:

$$E(g \mid G) = \frac{B(g \mid C)}{B(g \mid G)} \qquad [8]$$

where G is a set of highly expressed genes. Thus, E is dependent on the set G that can be chosen in different ways. In other words, the parameters of the model are the 61 codon fractions $X_{k,G}$ in the gene set G (see equation [6]).

Given this formal description of the CAI and the codon usage, the question is how we can use the genome-wide expression data to optimize the 61 parameters in the two models with respect to the prediction of expression levels.

Methods

Expression data

We give an overview of the expression data we used in this study in supplementary table 1. Briefly, we have combined different publicly available Affymetrix gene chip and SAGE datasets into one reference mRNA expression dataset, and two publicly available 2D-gel electrophoresis datasets into one reference protein abundance dataset (9-14). We have described this procedure, which helps to remove noise and errors from the data, previously (15). The codon composition of genes fundamentally affects the mechanism of protein translation; thus, the protein abundance data might contain more useful information than the mRNA expression data. On the other hand, the protein abundance data is available only for a very limited subset of 150 genes while there is a substantially larger amount of mRNA expression data (6071 genes). (For our calculations, we only considered those genes in the reference mRNA expression set that have an expression level of more than 0.5 copies/cell -- this is the case for 4270 genes. Smaller expression levels are too close to the resolution limits of the gene chips and therefore too noisy (see also caption of table 1)).

As described previously (15), we term the combination of a gene set (with G_{Prot} referring to the protein abundance and G_{mRNA} to the mRNA expression reference set) and an expression level or weight ($\boldsymbol{a_{Prot}}$ for protein and $\boldsymbol{a_{mRNA}}$ for mRNA abundance) "weighted population". Thus, three different weighted populations can be formed from our reference datasets: [G_{Prot}, $\boldsymbol{a_{Prot}}$], [G_{Prot}, $\boldsymbol{a_{mRNA}}$], and [G_{mRNA}, $\boldsymbol{a_{mRNA}}$]. ([G_{mRNA}, $\boldsymbol{a_{Prot}}$] is not meaningful since $\boldsymbol{a_{Prot}}$ is not defined on all genes in G_{mRNA}.) In the following we use all three populations for the parameterization of the CAI and the codon usage models.

Parameterization of the CAI and codon usage models with whole-genome expression data

Figure 1 schematically shows the procedure we used to parameterize the CAI and codon usage models with the expression data. We start by selecting one of the three populations mentioned above as an *evaluation set*. The evaluation set is later used to evaluate how well the CAI or codon usage model predicts actual expression levels. We also need to define a *parameterization set.* The parameterization set is the set of highly expressed genes G (see *Introduction*); it is used to calculate the parameters $w_k(G)$ for the CAI (see equation [3]) and the parameters $X_{k,G}$ for the codon usage (see equation [6]). To define the parameterization set, we choose one of the three populations and an expression level threshold T. We only include those genes of the population in the parameterization set whose expression level exceeds this threshold. With the parameters in hand, we are able to compute CAI and codon usage values for all genes in the evaluation set. We evaluate how well the CAI and codon usage models predict expression levels with two figures of merit: the Pearson correlation and the Spearman rank correlation[1].

We use the rank correlation as an additional diagnostic to the (linear) Pearson correlation because the relationship between CAI or codon usage values and expression levels is of a non-linear nature (see *Supplementary Material*).

We can iterate the procedure by changing the expression level threshold T and repeating the subsequent steps until we arrive at an optimal figure of merit. This gives us optimal parameters for the CAI and codon usage models.

Example of the CAI parameterization

Figure 2 shows a specific example of the parameterization of the CAI with $[G_{Prot}, \boldsymbol{a}_{\boldsymbol{Prot}}]$ as both the parameterization and evaluation population and illustrates how the figure of

[1] Given a set of abundance levels $\boldsymbol{a}$ in the evaluation set, and a vector of CAI or codon usage values ($\boldsymbol{C}$), we calculate the Pearson correlation as
corr(log($\boldsymbol{a}$),log($\boldsymbol{C}$))
and the rank correlation as
corr(rank($\boldsymbol{a}$),rank($\boldsymbol{C}$))

merit (Pearson correlation of the CAI values and the evaluation set) changes as a function of the expression level threshold T. When the threshold reaches T = 66,200 protein copies/cell the Pearson correlation reaches a maximum. At this point, there are only 21 genes in the parameterization set. The maximum correlation is slightly greater than the correlation between the CAI based on the original parameters by Sharp et al. (1) and the same evaluation set.

Linear model

In addition to the determination of the parameters for the CAI and codon usage models it is also possible to relate expression levels and codon composition of genes more directly.

The CAI formalism itself, slightly modified, suggests a multivariate linear model for doing this. Starting with equation [3], we can take the logarithm on both sides to obtain:

$$\log(CAI_g) = \sum_{k=1}^{61} X_{k,g} \log(w_k) \qquad [9]$$

If we introduce v_k _ log(w_k) and keep in mind that the log(CAI) is related to the logarithm of the gene expression levels, we can suggest the following linear model to predict the expression level a_g of a gene g:

$$y_g = v_0 + \sum_{k=1}^{61} X_{k,g} v_k \qquad [10]$$

with the residuals

$$\varepsilon_g = y_g - \log(a_g) \qquad [11]$$

In equation [9], y_g is the predicted expression level, the codon fractions $X_{k,g}$ are the predictor variables and v_0 … v_{61} the parameters. Note that we have introduced an intercept parameter v_0 in equation [10], for which there is no equivalent in equation [9]. We can then perform a standard multivariate linear regression to estimate the model parameters v_0 … v_{61} by minimizing the deviance:

$$d = \sum_g \varepsilon_g^2 \qquad [12]$$

Reducing the number of parameters in the linear model

One problem of this regression approach is obviously the large number of parameters. This may result in overfitting, even when the regression is applied to the largest population [G_{mRNA}, $\boldsymbol{a_{mRNA}}$], which contains 4270 data points.

We avoided this problem by deriving a linear model that consists of fewer parameters. This is done via a forward selection of parameters, adding one predictor variable at a time (25). A similar procedure has previously been used in finding significant promoter sequence motifs (18).

We start with a model of just one predictor variable (codon fraction X_k):

$$y_{k,g} = v_0 + v_k X_{k,g} \qquad [13]$$

which gives the residuals:

$$\varepsilon_{k,g} = \log(a_{mRNA,g}) - y_{k,g} \qquad [14]$$

and the deviance:

$$d_k = \sum_{g=1}^{4270} \varepsilon_{k,g}^2 \qquad [15]$$

Note that the deviance is dependent on the codon k. This allows us to find the codon that produces the smallest error and thus select the first predictor variable. We add this codon to a "model set" M.

Then we iterate this procedure. Given a model set M of codons with optimal parameter estimates, the linear model is:

$$y_{M,g} = v_0 + \sum_{m \in M} v_m X_{m,g} \quad [16]$$

This model gives the new residuals:

$$\varepsilon_{M,g} = \log(a_{mRNA,g}) - y_{M,g} \quad [17]$$

We then choose the next predictor variable by finding the codon k that minimizes:

$$\min_{k \notin M} \left(d_k = \sum_{g=1}^{4270} \left(\varepsilon_{M,g} - v_k X_k \right) \right) \quad [18]$$

This codon is then added to model set M, and we iterate the procedure described in equations [16]-[18]. Note that the interpretation of equation [18] is that the optimal predictor variable is orthogonal to the linear model of equation [16].

Significance of predictor variables

Each time we add a new predictor variable to the model, we need to check whether the corresponding parameter is significant. We can do this by observing the t statistic for a parameter estimate v_k; the ratio of a parameter estimate to its standard deviation follows a t-distribution and a P-value based on this distribution can be used for testing the hypothesis that $v_k = 0$. The t statistic and its corresponding P-values can be gathered from the standard output of a linear regression when performed in various statistical software packages (here, we used the publicly available R statistical computing environment, http://www.r-project.org/, as well as MATLAB for these computations).

To accept a predictor variable as significant we required that the P-value of the t statistic stay below $\alpha = 0.05$. Since we were choosing from several possible predictor variables at each step, a Bonferroni correction is necessary for this statistical test. This is equivalent to multiplying the P-value for a parameter with the number of remaining possible predictor variables. Given that there are already N_M parameters in the model set M, we have a choice of 61 - N_M remaining predictor variables, and the condition for significance is thus:

$$P' = (61 - N_M)P < \alpha \quad [19]$$

Results

Parameterization of the CAI and codon usage models

Table 1a shows the performance of the CAI and the codon usage with the original parameters in terms of the Pearson and rank correlation with the expression data. Here, the CAI parameters were taken from the original publication by Sharp et al. (1), which stem from 24 highly expressed genes. The situation is a little bit more complicated for the codon usage, in that previously the codon usage had not been explicitly used for the prediction of expression levels in yeast, but only in prokaryotes. However, to come up with a set of 'original' parameters, we computed them from the set of 128 ribosomal genes, following the recommendation of Karlin et al. who showed that, in yeast, ribosomal proteins exhibit the largest codon bias amongst all gene classes (4).

Table 1b generalizes the example shown in figure 1 by listing all possible evaluation and parameterization populations for both the CAI and the codon usage. Note that the parameters of the CAI and the codon usage are in each case dependent on the parameterization population and the expression level threshold *T*. (The threshold *T* defines the number of ORFs with expression levels greater than *T*). The table shows the maximum Pearson and rank correlations that can be achieved by varying *T*, the increase of the correlation compared with the original models ("Δ correlation"), and the size of the parameterization set at the maximum (rank) correlation, measured in number of ORFs.

A mixed picture emerges from this comprehensive collection of statistics. In many cases the new parameters improve the performance of the CAI and the codon usage (gray and black shaded squares in table 1b), but sometimes the performance is also slightly lower.

The codon usage models with the new parameters generally perform better than the model with the original parameters (Δ correlation is greater than 1% 6 out of 9 times for both the Pearson and rank correlations), whereas the improvements for the CAI are less

obvious (Δ correlation greater than 1% 3 out of 9 times for both the Pearson and rank correlations).

One important observation is that the parameterization sets for which we found optimal parameters are usually very small (on the order of 100 genes or less) for both the CAI and the codon usage. This is despite the fact that we used whole-genome information in our calculations. An extreme example is the codon usage with parameterization population [G_{Prot}, $\boldsymbol{a_{Prot}}$] and evaluation population [G_{Prot}, $\boldsymbol{a_{mRNA}}$]: here, the optimal parameterization set contains only one gene (the phosphopyruvate hydratase *ENO2*). This alone yields a rank correlation of 0.66 with the expression data.

Linear model

We fitted the linear model of equation [16] to the population [G_{mRNA}, $\boldsymbol{a_{mRNA}}$] according to the iterative procedure described in the *Methods* section. We tested models ranging from one to 61 codons (= predictor variables). The largest model for which all parameters were significant was a model with 20 codons. (The results for each model are shown in our supplementary material). These values of these 20 codon parameters are shown in figure 3. We have only used [G_{mRNA}, $\boldsymbol{a_{mRNA}}$] as the parameterization set because the other possible populations are too small (150 genes) relative to the possible number of parameters. When we used the reduced parameter procedure with [G_{prot}, $\boldsymbol{a_{prot}}$] or [G_{prot}, $\boldsymbol{a_{mRNA}}$] as the parameterization populations, we found that linear models with only two predictor variables are already superseding the critical *P*-value of 5% (see *Methods*), thus making them of little use for predicting expression levels.

The 20 codons that are significant predictor variables in the linear model for [G_{mRNA}, $\boldsymbol{a_{mRNA}}$] represent 13 different amino acids (see figure3). Of the seven remaining amino acids, 5 are underrepresented in highly expressed genes (Asp, His, Ile, Met and Tyr) while two of them are roughly equally represented in highly and lowly expressed proteins (15, 16). Four of the 20 chosen predictor variables (= codon compositions) are negatively correlated with expression levels. The parameters of the linear model and corresponding codons (= predictor variables) are discussed in more detail in the section "preferential

codons in yeast" below. Details of the regression results (parameters, P-values etc.) can be found in our *Supplementary Material*.

The bottom of table 1b shows the performance of the linear model compared with the CAI and codon usage. There is no possible comparison to a set of original parameters, as in the case of the CAI and the codon usage. Instead, we compared the performance of the linear model with the performance of the original CAI and codon usage models on the same evaluation sets. The left half of the "Δ correlation" column in table 1b refers to the difference with the CAI correlation, whereas the right half gives the difference with the codon usage correlation. (There are three possible choices for the evaluation set.) It is clear from the results that the best performance is obtained when the parameterization and evaluation populations are both [G_{mRNA}, $\boldsymbol{a_{mRNA}}$]. (This should be expected, given that the model parameters were optimized on this set.)

When [G_{mRNA}, $\boldsymbol{a_{mRNA}}$] is both the parameterization and evaluation population, the Pearson correlation of the linear model with the expression data is 0.75. This is slightly higher than the best Pearson correlations for the CAI and CU models. (The CAI has a maximum Pearson correlation of 0.72, while the CU has a maximum Pearson correlation of 0.71.) In terms of the rank correlation, the best CU model is somewhat better than the linear model (0.60 vs. 0.56), while the CAI performs worse than both of the other methods (0.46).

Preferential codons in yeast

As mentioned in the beginning, it is important for heterologous gene expression to encode proteins with sequences that yield optimal expression. A good rule of thumb for finding such an optimal sequence is to choose codons that are most frequent in highly expressed genes. The CAI model provides an explicit way of finding such codons; the most frequent codons simply have the highest relative adaptiveness values, and sequences with higher CAIs are preferred over those with lower CAIs. The codon usage formalism does not explicitly use relative adaptiveness values, but they can be easily calculated with equation [1] from the parameterization sets that yield optimal codon usage parameters. A

third possibility is to interpret the parameters of the linear regression with respect to which codons are more preferred. (This is of course only possible for those codons that are predictor variables in the linear model.)

Figure 3 shows the relative adaptiveness values for the CAI and codon usage (CU) -- when the parameterization and evaluation populations are both [G_{Prot}, $\boldsymbol{a}_{\boldsymbol{Prot}}$] with the Pearson correlation as the figure of merit -- together with the parameter values of the linear regression (LM) with [G_{mRNA}, $\boldsymbol{a}_{\boldsymbol{mRNA}}$]. For comparison, we also show the relative adaptiveness values for the genome as a whole. Codons with relative adaptiveness values of 100% (= preferential codons) are shown in black. It is evident that both the CAI and the codon usage give the same preferential codons.

The relative adaptiveness values for the CAI are computed from the 21 most abundant proteins in $\boldsymbol{a}_{\boldsymbol{Prot}}$, whereas the codon usage values stem from the 4 most abundant proteins. Note that the preferential codons for both the CAI and the codon usage stay the same regardless of which parameterization and evaluation sets we choose (with the Pearson correlation as the figure of merit). The only exception is when we choose [G_{mRNA}, $\boldsymbol{a}_{\boldsymbol{mRNA}}$] as both the parameterization and evaluation set for the codon usage. In that case, the optimal parameterization set becomes relatively large (253 ORFs) such that several of the preferential codons are the same as the ones for the genome as a whole.

The parameters of the linear model are shown in the third column for each codon in figure 3. Not that the parameters v_k give the expected change of expression level for an increase in the composition of the corresponding codon k, given that the composition of the other codons in the model stays the same:

$$\frac{\partial y_g}{\partial X_{g,k}} = v_k \qquad [20]$$

One would expect the regression parameters to roughly correlate with the relative adaptiveness values of the CAI and codon usage. Because the number of parameters in the linear model is less than the total number of codons, this comparison is only possible for synonymous codons of seven amino acids (see figure 3).

Contrary to our expectation, the rank order of the regression parameters was different than that of the relative adaptiveness values of the CAI and codon usage for three of these seven amino acids (Val, Cys and Arg). One (non-biological) explanation for this different order might be the sensitivity of the parameters. This is in fact the case for Val and Cys where the 95% confidence intervals of the parameter values overlap (see *Supplementary Material*). However, parameter sensitivity does not explain the different codon order for Arginine; the codon CGT has a much higher parameter value than the codon AGA (9.7 as opposed to 4.7), contrary to the ranking of relative adaptiveness values (see figure 3).

In contrast to the linear model parameters, the relative adaptiveness values describe the global enrichment of a codon in highly expressed genes with no restrictions on the compositions of the other codons. (This is confirmed by the fact that the Pearson correlation between the logarithms of a_{mRNA} and the codon composition of AGA is larger than that between a_{mRNA} and CGT). Thus, in the case of arginine, one explanation for the discrepancy between the linear model and the CAI/codon usage might be that yeast cells preferentially use AGA codons for Arginine in highly expressed genes (explaining the CAI value), but that the supply of the corresponding tRNA is already strongly exhausted for fast growing cells. Thus, to achieve additional translation of Arginine at high rates, the cell might need to use the supply of another tRNA for Arginine (explaining the higher regression parameter for AGA). Note that the tRNA gene copy number is 11 for the AGA codon and 6 for the CGT codon (the highest and second highest among all Arginine codons). This way, the cell would make optimal use of the supply of Arginine tRNAs when it is already growing fast.

Discussion

Quantitative versus qualitative, genome-wide versus few genes

The CAI and codon usage models are originally based on somewhat qualitative assumptions about the expression levels of relatively few genes. This was our motivation for using quantitative, genome-wide expression data to recalculate optimal model

parameters. These new parameters sometimes lead to a slightly better correlation of the codon-based expression models with expression data according to several measures, although the improvements are relatively marginal and the results are mixed.

Small parameterization sets are sufficient

Furthermore, the parameterization sets that yielded optimal parameters for the CAI and codon usage are often very small compared to the number of genes in the genome -- very much in the same way that the original parameterization sets were small (see table 1). Thus, very few highly expressed genes often seem to be sufficient to describe the overall codon bias in yeast. This shows that the original procedures for determining the parameters of the CAI and codon usage were indeed quite prescient. The CAI and codon usage models are relatively insensitive to the exact choice of highly expressed genes.

One explanation for this observation might be that although the optimal parameterization sets are small compared to the size of the genome, their share of the overall number of transcripts and protein copies in the cell is much larger; they may in fact dominate the overall codon composition of transcripts and proteins (16). This situation can be compared with the way a financial market index, composed of very few stocks with very high market capitalization, can be a very good approximation for the value of a total market, which consists of perhaps thousands of individual stocks.

Thus, to obtain robust parameters for the CAI and codon usage models, it often seems sufficient to infer them from rather qualitative information about gene expression levels. For instance, it may be enough to infer from information about biological function whether a group of genes is highly expressed. Note that, using our parameterization procedure, we achieved a Pearson correlation of 0.72 between the codon usage model and the expression data (when both the evaluation and parameterization population are $[G_{mRNA}, \boldsymbol{a_{mRNA}}]$, see table 1b). This is only a marginal improvement over the original parameters (Pearson correlation 0.71, see table 1a) that were derived from the codon composition of the 128 ribosomal proteins in yeast.

Comparison of the CAI, codon usage and linear models

In contrast to the linear model and the codon usage, the parameters of the CAI are normalized by synonymous codon usage, a constraint that is not present in the other two models. It is therefore remarkable that the CAI model (given the best parameterization set) usually performs as well as the other two models. The only exception from this general rule is perhaps the relatively low rank correlation of the CAI with [G_{mRNA}, $\boldsymbol{a}_{\boldsymbol{mRNA}}$], which is only 0.49 under the best circumstances (compared with 0.60 for the codon usage and 0.56 for the linear model).

The linear model achieves the highest Pearson correlation (0.75) with [G_{mRNA}, $\boldsymbol{a}_{\boldsymbol{mRNA}}$], while the comparable values for the CAI and codon usage are slightly lower (0.72 and 0.71).

Can the models be improved?

The main motivation of our study was the question whether it would be possible to improve on existing and commonly used codon-based models for predicting expression levels. The results showed that the original models are relatively robust to the exact way they are parameterized. However, one could argue that these models could still be improved if other protein properties were included as additional features in the prediction.

We have explicitly tested whether one protein property, namely protein length, can aid in improving the prediction performance. It has previously been observed that longer proteins often tend to be less highly expressed than shorter ones (16, 19). For instance, in the linear regression model one could explicitly consider protein length by replacing the codon fractions X_k with the number of codons (equation [16]). However, we found that this severely decreases the correlation between the model predictions and actual expression data (data not shown).

Codon composition is often the most strong predictor of expression levels

Pavesi (20) proposed a model for predicting expression levels based on several different protein properties (the CAI, the codon bias index, an entropy score relating to synonymous codon usage, a TATA-box score and a pyrimidine bias index) (21). He showed in a regression analysis that the two significant parameters of the model were the CAI and the entropy score, both measures relating to synonymous codon usage. Pavesi reported a Pearson correlation of 0.76 with a select set of 621 SAGE expression levels.

Linear model

As an alternative to the CAI and codon usage models, we have proposed a simple linear model that relates codon fractions and expression levels of genes. An advantage of the linear model is that, unlike the numerical values from the CAI and the codon usage, the predicted expression levels have the same dimension (copies/cell) as the actual expression levels and are directly comparable with them. The linear model predicts an expression level of 1.7 copies/cell for transcripts from sequences with average codon fractions; this is equal to the average expression level in a_{mRNA}. (This follows from equation [11] and the fact that the average residual in the model is equal to zero.)

We have suggested a natural, intuitive justification for the linear model, based on the CAI formalism. Of course, this does not exclude the possibility that there are better alternatives. From a mathematical standpoint, the linear regression is relatively simple and involves much less complex computations than non-linear regressions.

Applications

Overall, it seems justified to use the CAI, codon usage or related measures as "rules of thumb" in a variety of applications such as heterologous gene expression, either based on the original parameters or on our newly optimized ones. For the annotation of genomes, all three models seem to be useful, however, they should of course only be used in conjunction with other gene-finding criteria (22).

The 20-parameter linear model allows us to compare the codon parameters for seven amino acids. Surprisingly, the linear model parameters suggest a different rank order for the codons of the amino acid Arginine. An explanation might be that fast growing yeast cells have already exhausted the supply of the most optimal codon, and thus have to make use of the second best codon.

General issues of data quality

The value of the codon-based expression indicators can perhaps be appreciated by comparing them to the correlation of mRNA and protein abundance data in general. The correlation for the two populations [G_{Prot}, $\boldsymbol{a_{mRNA}}$] and [G_{Prot}, $\boldsymbol{a_{Prot}}$] is 0.67, well within the range of the correlations in table 1 (13-15). One interpretation of this is that the codon-based expression indicators are actually just as good as mRNA expression data in the prediction of protein abundance levels.

Of course, the codon-based expression indicators yield static values, whereas gene expression is of course a dynamic process, with very different expression levels under different conditions. The expression data that we used in this study stems from experiments under very similar conditions, that is, yeast cells in vegetative growth on rich media (9-12). Thus, the prediction of expression levels based on codon composition should work best for these physiological situations, but might work less well for others. Coghlan et al. have pointed to the example of *ENO1* and *ENO2*, which both exhibit strong codon biases -- the former is repressed by high glucose concentrations whereas the latter is strongly induced (19). In general, the regulation of translation might be less flexible than the regulation of transcription because the abundance of charged tRNAs cannot be changed as flexibly as the abundance of transcription factors (there are 33 cognate tRNAs in yeast, but perhaps hundreds of transcription factors (23, 24).

Of course, there are many limitations of the expression data itself that might confound the relationship between expression levels and codon composition. The 2D-gel data is subject to many biophysical and biochemical constraints (13, 14, 17). The situation is

somewhat better for the mRNA expression data, where we have more data resources that we combined in this study.

Supporting website

Additional data relating to our analysis is available at:
http://bioinfo.mbb.yale.edu/expression/codons

Acknowledgements

We thank M. Seringhaus and H. Bussemaker for helpful discussions. MG acknowledges support from NIH grant P20 LM07253-01.

Figures and tables

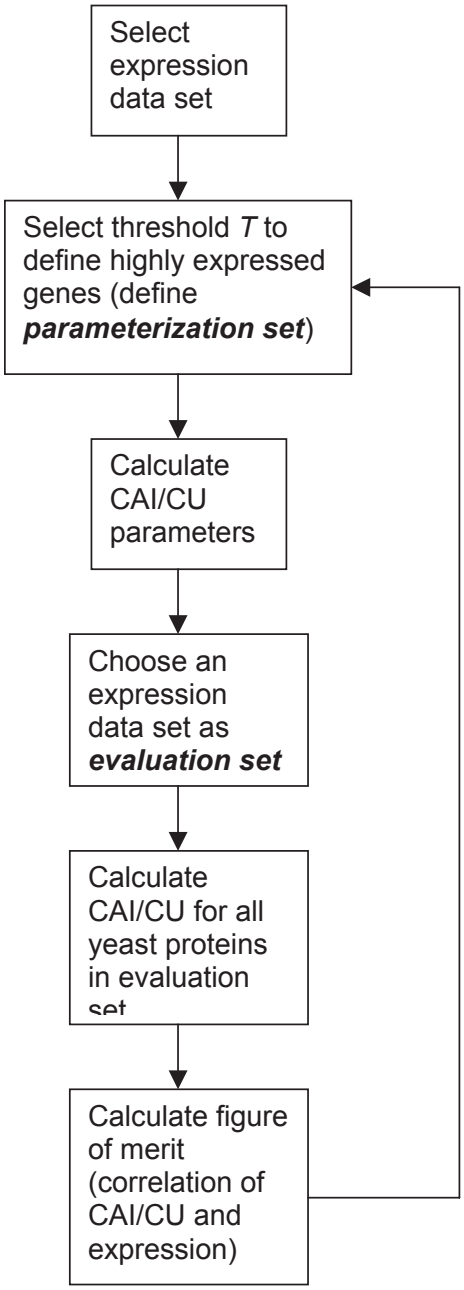

Figure 1

Our general procedure for the parameterization of the CAI and codon usage models. We first choose an expression data set and an arbitrary expression level threshold *T* to differentiate highly from lowly expressed genes. The highly expressed genes with

expression levels greater than *T* define the *parameterization set*. Based on this we calculate new model parameters. Finally, to evaluate the performance of the models, we choose another expression data set (we term this the *evaluation set*): we calculate the CAI and codon usage values for all genes in the evaluation set and then measure the correlation between the model values and the actual expression levels as a figure of merit.

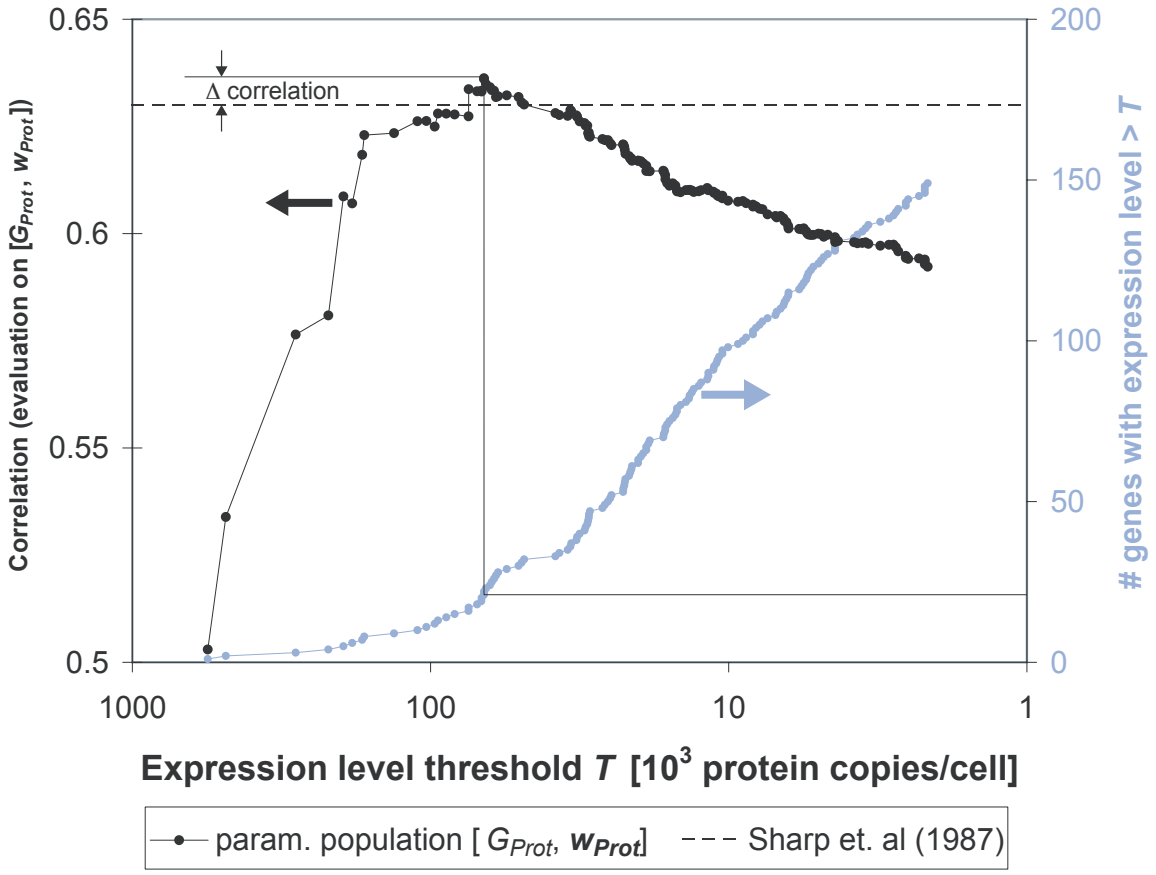

Figure 2

An example of the parameterization of the CAI with expression data. Here, we use [G_{Prot}, $\boldsymbol{a_{Prot}}$] for both the parameterization and the evaluation steps. The Pearson correlation of the CAI with the evaluation set (left ordinate) is shown as a function of the expression level threshold T, which defines the parameterization set of highly expressed genes. The right ordinate shows the number of genes in the parameterization set for a given threshold T. At T = 66,200 proteins/cell, the Pearson correlation reaches a maximum. This correlation is slightly higher than the correlation of the original CAI model with the evaluation set (dashed line).

First position	Second position																Third position
	T					C				A				G			
T	20	4		**100**	Phe	**100**	**100**	**3.2**	**100**	10	0		**100**	Tyr	**100**	**100**	**5.2**	**100**	Cys	T
	100	**100**	**4.7**	68		84	79		60	**100**	**100**		76		9	8	**8.5**	61		C
	15	13		99	Leu	8	2		81	-	-		-	*Stop*	-	-		-	*Stop*	A
	100	**100**	**5.9**	**100**		3	0		38	-	-		-		**100**	**100**		**100**	Trp	G
C	2	1		47		15	7		76	20	16		**100**	His	9	8	**9.7**	30	Arg	T
	1	0		21		2	1	**-8.8**	39	**100**	**100**		56		0	0		13		C
	6	4		51		**100**	**100**	**7.5**	**100**	**100**	**100**	**3.6**	**100**	Gln	0	0		15		A
	2	0		40		1	0		31	7	2		46		0	0		9		G
A	90	73		**100**	Ile	88	85		**100**	11	6	**-7.3**	**100**	Asn	7	2		62	Ser	T
	100	**100**		56		**100**	**100**		62	**100**	**100**	**5.5**	68		12	4		43		C
	2	1		61		8	1		90	16	7		**100**	Lys	**100**	**100**	**4.7**	**100**	Arg	A
	100	**100**		**100**	Met	2	1		41	**100**	**100**	**7.7**	72		1	0	**-7.5**	45		G
G	**100**	**100**	**4.9**	**100**	Val	**100**	**100**	**8.9**	**100**	52	38		**100**	Asp	**100**	**100**	**9.0**	**100**	Gly	T
	96	95	**7.5**	52		38	31		61	**100**	**100**		53		6	3		44		C
	3	1		56		2	2	**-6.6**	81	**100**	**100**		**100**	Glu	2	0	**-4.6**	50		A
	9	3		51		1	0		31	7	3		43		6	3		27		G

CAI CU LM Genome

Figure 3

Shows which codons are common in highly expressed genes. There are four columns for each codon. The first two columns show the relative adaptiveness values for the CAI and codon usage (CU) according to equation [1]. The third column shows the regression parameters of the linear model (LM). Note that there are only 20 values because the model contains only 20 codons as predictor variables. The fourth columns shows the relative adaptiveness values for the genome as a whole. The relative adaptiveness values are normalized to 100 for the most frequent synonymous codons. The regression parameters are not normalized.

a

Model	Parameters according to …	Evaluation set	Pearson correlation	Rank correlation	# ORFs in parameteri-zation set
CAI	Sharp et al.(1987)	$[G_{Prot}, a_{Prot}]$	0.63	0.61	24
		$[G_{Prot}, a_{mRNA}]$	0.70	0.70	
		$[G_{mRNA}, a_{mRNA}]$	0.69	0.43	
codon usage (CU)	Karlin et al. (1998)	$[G_{Prot}, a_{Prot}]$	0.63	0.61	128
		$[G_{Prot}, a_{mRNA}]$	0.67	0.65	
		$[G_{mRNA}, a_{mRNA}]$	0.71	0.54	

b

Model	Parameteri-zation set chosen from …	Evaluation set	Maximum Pearson correlation	Δ Pearson correlation		# ORFs in parameteri-zation set at maximum correlation	Maximum rank correlation	Δ rank correlation		# ORFs in parameteri-zation set at maximum correlation
				rel. to CAI	rel. to CU			rel. to CAI	rel. to CU	
CAI	$[G_{Prot}, a_{Prot}]$	$[G_{Prot}, a_{Prot}]$	0.64	0.6%	-	21	0.62	0.8%	-	21
		$[G_{Prot}, a_{mRNA}]$	0.70	0.0%	-	22	0.71	0.5%	-	36
		$[G_{mRNA}, a_{mRNA}]$	0.71	2.0%	-	14	0.47	4.0%	-	12
	$[G_{Prot}, a_{mRNA}]$	$[G_{Prot}, a_{Prot}]$	0.63	-0.4%	-	7	0.61	-0.2%	-	47
		$[G_{Prot}, a_{mRNA}]$	0.70	0.0%	-	37	0.71	0.5%	-	46
		$[G_{mRNA}, a_{mRNA}]$	0.72	3.0%	-	15	0.49	6.0%	-	15
	$[G_{mRNA}, a_{mRNA}]$	$[G_{Prot}, a_{Prot}]$	0.63	-0.4%	-	27	0.61	-0.2%	-	120
		$[G_{Prot}, a_{mRNA}]$	0.70	0.0%	-	106	0.71	0.5%	-	103
		$[G_{mRNA}, a_{mRNA}]$	0.70	1.0%	-	23	0.46	3.0%	-	220
codon usage (CU)	$[G_{Prot}, a_{Prot}]$	$[G_{Prot}, a_{Prot}]$	0.69	-	5.9%	4	0.63	-	2.4%	8
		$[G_{Prot}, a_{mRNA}]$	0.70	-	3.1%	7	0.66	-	0.5%	1
		$[G_{mRNA}, a_{mRNA}]$	0.70	-	-1.0%	62	0.59	-	5.0%	150
	$[G_{Prot}, a_{mRNA}]$	$[G_{Prot}, a_{Prot}]$	0.68	-	4.9%	6	0.63	-	2.4%	6
		$[G_{Prot}, a_{mRNA}]$	0.71	-	4.1%	3	0.66	-	0.5%	12
		$[G_{mRNA}, a_{mRNA}]$	0.71	-	0.0%	44	0.59	-	5.0%	149
	$[G_{mRNA}, a_{mRNA}]$	$[G_{Prot}, a_{Prot}]$	0.68	-	4.9%	5	0.63	-	2.4%	5
		$[G_{Prot}, a_{mRNA}]$	0.71	-	4.1%	3	0.66	-	0.5%	5
		$[G_{mRNA}, a_{mRNA}]$	0.71	-	0.0%	253	0.60	-	6.0%	983
Linear regression	$[G_{mRNA}, a_{mRNA}]$	$[G_{Prot}, a_{Prot}]$	0.65	1.6%	1.9%	-	0.60	-1.2%	-0.6%	-
		$[G_{Prot}, a_{mRNA}]$	0.65	-5.0%	-1.9%	-	0.41	-29.5%	-24.5%	-
		$[G_{mRNA}, a_{mRNA}]$	0.75	6.9%	4.9%	-	0.56	13.0%	2.0%	-

Table 1

Table 1a shows the Pearson and rank correlation of the original CAI and codon usage models with various evaluation sets of expression data. The last column of the table

shows how many genes were used to calculate the original parameters (given in number of ORFs).

Table 1b shows the Pearson and rank correlations of the CAI and codon usage models based on the new parameters. Altogether, there are nine possible combinations of parameterization and evaluation sets (2nd and 3rd column) for both models. The fourth column shows the Pearson correlation for the best set of parameters that could be found with our procedure. The fifth columns compares this value with the correlation achieved by the original models. The seventh column shows how many genes were present in the parameterization set that yielded that maximum correlation. The remaining columns give the same statistics for the rank correlation.

The bottom of the table shows similar statistics for the linear model.

References

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

2. Ikemura, T. (1981) Correlation between the abundance of Escherichia coli transfer RNAs and the occurrence of the respective codons in its protein genes: a proposal for a synonymous codon choice that is optimal for the E. coli translational system. *J Mol Biol*, **151**, 389-409.

3. Karlin, S., Mrazek, J. and Campbell, A. M. (1998) Codon usages in different gene classes of the Escherichia coli genome. *Mol Microbiol*, **29**, 1341-55.

4. Karlin, S., Campbell, A. M. and Mrazek, J. (1998) Comparative DNA analysis across diverse genomes. *Annu Rev Genet*, **32**, 185-225.

5. Sharp, P. M. and Matassi, G. (1994) Codon usage and genome evolution. *Curr Opin Genet Dev*, **4**, 851-60.

6. Bennetzen, J. L. and Hall, B. D. (1982) Codon selection in yeast. *J Biol Chem*, **257**, 3026-31.

7. Karlin, S., Mrazek, J., Campbell, A. and Kaiser, D. (2001) Characterizations of highly expressed genes of four fast-growing bacteria. *J Bacteriol*, **183**, 5025-40.

8. Karlin, S. and Mrazek, J. (2000) Predicted highly expressed genes of diverse prokaryotic genomes. *J Bacteriol*, **182**, 5238-50.

9. Holstege, F. C., Jennings, E. G., Wyrick, J. J., Lee, T. I., Hengartner, C. J., Green, M. R., Golub, T. R., Lander, E. S. and Young, R. A. (1998) Dissecting the regulatory circuitry of a eukaryotic genome. *Cell*, **95**, 717-28.

10. Jelinsky, S. A. and Samson, L. D. (1999) Global response of Saccharomyces cerevisiae to an alkylating agent. *Proc Natl Acad Sci U S A*, **96**, 1486-91.

11. Roth, F. P., Hughes, J. D., Estep, P. W. and Church, G. M. (1998) Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation. *Nat Biotechnol*, **16**, 939-45.

12. Velculescu, V. E., Zhang, L., Zhou, W., Vogelstein, J., Basrai, M. A., Bassett, D. E., Jr., Hieter, P., Vogelstein, B. and Kinzler, K. W. (1997) Characterization of the yeast transcriptome. *Cell*, **88**, 243-51.

13. Gygi, S. P., Rochon, Y., Franza, B. R. and Aebersold, R. (1999) Correlation between protein and mRNA abundance in yeast. *Mol Cell Biol*, **19**, 1720-30.

14. Futcher, B., Latter, G. I., Monardo, P., McLaughlin, C. S. and Garrels, J. I. (1999) A sampling of the yeast proteome. *Mol Cell Biol*, **19**, 7357-68.

15. Greenbaum, D., Jansen, R. and Gerstein, M. (2002) Analysis of mRNA expression and protein abundance data: an approach for the comparison of the enrichment of features in the cellular population of proteins and transcripts. *Bioinformatics*, **18**, 585-96.

16. Jansen, R. and Gerstein, M. (2000) Analysis of the yeast transcriptome with structural and functional categories: characterizing highly expressed proteins. *Nucleic Acids Res*, **28**, 1481-8.

17. Gygi, S.P., Corthals, G. L., Zhang, Y., Rochon, Y., Aebersold, R. (2000) Evaluation of two-dimensional gel electrophoresis-based proteome analysis technology. *Proc Natl Acad Sci U S A*, **97**, 9390-5.

18. Bussemaker, H.J., Li, H., Siggia, E.D. (2001) Regulatory element detection using correlation with expression *Nat Genet*, **27**, 167-71.

19. Coghlan, A., Wolfe, K.H. (2000) Relationship of codon bias to mRNA concentration and protein length in Saccharomyces cerevisiae. *Yeast*, **16**, 1131-45.

20. Pavesi, A. (1999) Relationships between transcriptional and translational control of gene expression in Saccharomyces cerevisiae: a multiple regression analysis. *J Mol Evol*, **48**, 133-41.

21. Konopka, A. (1984) Is the information content of DNA evolutionarily significant? *J Theor Biol* **107**, 697–705.

22. Kumar, A., Harrison, P.M., Cheung, K.H., Lan, N., Echols, N., Bertone, P., Miller, P., Gerstein, M., Snyder, M. (2002) An integrated approach for finding overlooked genes in yeast. *Nat Biotechnol*, **20**, 58-63.

23. Horak, S.E., Luscombe, N.M., Qian, J., Bertone, P., Piccirrillo, S., Gerstein, M., Snyder, M. (2002) Complex transcriptional circuitry at the G1/S transition in Saccharomyces cerevisiae. *Genes Dev*, **16**, 3017-33.

24. Lee, T.I., Rinaldi, N.J., Robert, F., Odom, D.T., Bar-Joseph, Z., Gerber, G.K., Hannett, N.M., Harbison, C.T., Thompson, C.M., Simon, I. et al. (2002) Transcriptional regulatory networks in Saccharomyces cerevisiae. *Science*, **298**, 763-4.

25. Dobson, J.D., Applied multivariate data analysis, volume I: regression and experimental design, Springer (1999).