Correlation between Expression Level of Gene and Codon Usage

G P S Raghava[1,2*], Da Jung Hwang[1], and Joon Hee Han[1]

[1]Department of Computer Science and Engineering, Pohang Un iversity of Science and Technology, San 31 Hyo-Ja Dong, Pohang 790-784, Republic of Korea

[2]Bioinformatics Centre, Institute of Microbial Technology, Sector 39A, Chandigarh -160036, India

[*]To whom correspondence should be addressed. E-mail: raghava@imtech.res.in

Abstract

In this study, we analyzed the gene expression data of *Saccharomyces cerevisiae* obtained from Holstege et al. 1998 to understand the relationship between expression level and nucleotide sequence of a gene. First, the correlation between gene expression and percent composition of each type of nucleotide was computed. It was observed that nucleotide '**G**' and '**C**' show positive correlation ($r \geq 0.15$), '**A**' shows negative correlation ($r \approx -0.21$) and 'T' shows no correlation ($r \approx 0.00$) with gene expression. It was also found that 'G+C' rich genes express more in comparison to 'A+T' rich genes. We observed the inverse correlation between composition of a nucleotide at genome level and level of gene expression. Then we computed the correlation between dinucleotides (e.g. AA, AT, GC) composition and gene expression and observed a wide variation in correlation (from r = -0.45 for AT to r = 0.35 for GT). The dinucleotides which contain 'T' have wide range of correlation with gene expression. For example, GT and CT have high positive correlation and AT have high negative correlation. We also computed the correlation between trinucleotides (or codon) composition and gene expression and again observed wide range of correlation (from r = -0.45 for ATA r = 0.45 for GGT). However, the major codons of a large number of amino acids show positive correlation with expression level, but there are a few amino acids whose major codons show negative correlation with expression level. These observations clearly indicate the relationship between nucleotides composition and expression level. We also demonstrate that codon composition can be used to predict the expression of gene in a given condition. Software has been developed for calculating correlation between expression of gene and codon usage.

Introduction

The gene expression is a complex and context dependent phenomena which play a major role in function, evolution and survival of an organism (1, 2). Fortunately, we have powerful techniques like DNA microarray that allows monitoring of expression level of several thousands of genes

simultaneously. This massive expression data provides us unique opportunities to detect co-regulatory genes, clustering of genes, evolutionary genes and gene network etc. Due to this technology we have expression data of a large number of organisms in various conditions, which posed a major challenge to the bioinformaticians to deduce the relationship between gene expression and nucleotide sequence of genome. It has been shown in past that GC rich genes have different expression than AT rich genes in genomes. Though there are indirect studies, which indicates relation between sequence but there is no direct study on correlation between sequence and expression (3, 4). First time we made an systematic attempt to deduce the relation between nucleotide sequence and expression of gene. We compute correlation between expression and various type of compositions of gene (e.g., single, di-nucleotide and tri-nucleotide compositions).

Materials & Methods

Dataset

There are number of databases which provides maintain microarray data. The selection of appropriate data for analysis is one of the important issues for performing this type of study. In this study we perform analysis on yeast, which is studied in detail by number of groups, as well as it is fully annotated (5). In this study, we analyzed the gene expression data of *Saccharomyces cerevisiae* obtained from Holstege et al. 1998 to understand the relationship between expression level and nucleotide sequence of a gene (http://www.wi.mit.edu/young/expression.html/). We selected this dataset for our study because its results are obtained from careful averaging of many experiments, as well as this dataset was widely used by researchers (6). Our dataset contains 4807 genes; whose nucleotide sequences are available in Saccharomyces Genome Database (SGD) at http://www.yeastgenome.org/.

Nucleotide Compositions: In this study we compute the correlation between expression and different type of nucleotide compositions. The type of nucleotide compositions basically percent composition, we calculate in this study includes; i) single nucleotide; ii) dinucleotides and iii) trinucleotides. Following is the procedure we adopt to calculate these compositions.

Percent Composition of Single Nucleotides: The information of a gene can be encapsulated in a vector of 4 dimensions using single nucleotide composition of the gene. The composition was used as input in this study, which provides the global information of gene features in the form of fixed length vector. We used the following procedure to calculate composition, lets a gene has N nucleotides (NT), and there are 4 types of nucleotide A, T, G, C. Then we can calculate the percent composition of nucleotide A (A_p) using following equation.

$$A_p = \frac{\sum_{i=1}^{N} NT_i}{N} \times 100 \qquad (1)$$

(if NT_i = A then NT_i = 1, otherwise NT_i = 0)

Similarly, we can calculate the percent composition of other three nucleotides. In this way a gene can be expressed by four numbers (percent composition of nucleotides).

Percent Composition of Dinucleotides: In case of single nucleotide there was no information about order where as in dinucleotides we also consider the local order of two nucleotides. As there are 4 nucleotides so they can make 16 possible dinucleotides like AT, AG, AC, AA, TA, TG, TC, TT etc. We can calculate the percent composition of a dinucleotide AT (AT_p) using following equation

$$AT_p = \frac{\sum_{i=1}^{N-1} NT_i NT_{i+1}}{(N-1)} \times 100 \qquad (2)$$

(if NT_i = A and NT_{i+1} = T then NT_i NT_{i+1} = 1, otherwise NT_i NT_{i+1} = 0)

Similarly, one can calculate the percent composition of remaining dinucleotides. In this way a gene can be expressed by 16 numbers (percent composition of dinucleotides)

The above procedure is to compute the dinucleotide composition in overlapping mode where all possible dinucleotides have been calculated. In this study we also compute the dinucleotide composition in two possible frames using following equations (example for computing percent composition of AT (AT_{p1} and AT_{p2}).

Frame 1

$$AT_{p1} = \frac{\sum_{i=1}^{\frac{N-1}{2}} NT_{2i-1} NT_{2i}}{\frac{N}{2}} \times 100 \qquad (3)$$

(if NT_{2i-1} = A and NT_{2i} = T then NT_{2i-1} NT_{2i} = 1, otherwise NT_{2i-1} NT_{2i} = 0)

Frame 2

$$AT_{p2} = \frac{\sum_{i=1}^{\frac{N-1}{2}} NT_{2i} NT_{2i+1}}{\frac{N-1}{2}} \times 100 \qquad (4)$$

(if NT_{2i} = A and N_{2i+1} = T then NT_{2i} NT_{2i+1} = 1, otherwise NT_{2i} NT_{2i+1} = 0)

Similarly, one can calculate the percent composition of remaining dinucleotides

Percent Composition of Trinucleotides or Codons: In this case we consider three continuous nucleotides, this is similar to codon. The total number of possible trinucleotides made by four nucleotides will be 64 like AGC, AAA, AAT, AAG etc. We can calculate the percent composition of trinucleotides AGC_p using following equation

$$AGC_p = \frac{\sum_{i=1}^{N-2} NT_i NT_{i+1} NT_{i+2}}{N-2} \times 100 \qquad (5)$$

(if NT_i = A , NT_{i+1} = G and NT_{i+2} = C then NT_i NT_{i+1} NT_{i+2} = 1, otherwise NT_i NT_{i+1} NT_{i+2} = 0)

Similarly, one can calculate the percent composition of remaining trinucleotides. In this way a gene can be expressed by 64 numbers (percent composition of trinucleotides or codons)

In this study we also compute percent composition of codons in all three possible fames using equations given below. Following examples shows how to compute percent composition of codon AGC in three frames (AGC_{p1}, AGC_{p2}, AGC_{p3})

Frame 1

$$AGC_{p1} = \frac{\sum_{i=1}^{\frac{N}{3}} NT_{3i-2} NT_{3i-1} NT_{3i}}{\frac{N}{3}} \times 100 \qquad (6)$$

(if NT_{3i-2} = A , NT_{3i-1} = G and NT_{3i} = C then NT_{3i-2} NT_{3i-1} NT_{3i} = 1, otherwise NT_{3i-2} NT_{3i-1} NT_{3i} = 0)

Frame 2

$$AGC_{p2} = \frac{\sum_{i=1}^{\frac{N-1}{3}} NT_{3i-1} NT_{3i} NT_{3i+1}}{\frac{N-1}{3}} \times 100 \qquad (7)$$

(if NT_{3i-1} = A , NT_{3i} = G and NT_{3i+1} = C then NT_{3i-1} NT_{3i} NT_{3i+1} = 1, otherwise NT_{3i-1} NT_{3i} NT_{3i+1} = 0)

Frame 3

$$AGC_{p3} = \frac{\sum_{i=1}^{\frac{N-2}{3}} NT_{3i} NT_{3i+1} NT_{3i+2}}{\frac{N-2}{3}} \times 100 \qquad (8)$$

(if NT_{3i} = A , NT_{3i+1} = G and NT_{3i+2} = C then NT_{3i} NT_{3i+1} NT_{3i+2} = 1, otherwise NT_{3i} NT_{3i+1} NT_{3i+2} = 0)

Similarly, one can calculate the percent composition of remaining trinucleotides.

Correlation between Gene Expression and Nucleotide Composition: First, we compute the percent composition of single nucleotide, dinucleotides, and trinucleotides corresponding to each gene in our data set of 4807 genes. Then we compute Pearson's correlation coefficient (r), which is the ratio of the covariance between the percent composition of nucleotide and gene expression level to the product of the standard deviations in the two.

$$r = \frac{\Sigma XY - \frac{\Sigma X \Sigma Y}{N}}{\sqrt{\left(\Sigma X^2 - \frac{(\Sigma X)^2}{N}\right)\left(\Sigma Y^2 - \frac{(\Sigma Y)^2}{N}\right)}} \qquad (9)$$

where, X and Y are nucleotide composition and expression level of gene respectively. N is the total number of genes in the data set which is 4807 in our case

Results

Single Nucleotide Composition: First we computed the single nucleotide composition of each gene in our dataset using equation 1 and then we calculated the correlation between expression and composition using equation 9. As shown in Table 1, nucleotide 'A' shows high negative correlation, 'G' and 'C' show positive correlation and 'T' do not exhibit any correlation with level of gene expression. We also computed the correlation between gene expression and total content of GC (percent of G + C) and observed a positive correlation. Similarly correlation was computed between gene expression and content of AT (percent of A + T) and negative correlation was observed. Thus GC rich genes have higher expression then AT rich genes in general, in yeast. It is interesting that correlation is inversely proportional to overall percent composition of nucleotide. As shown in Table 1, 'C' have minimum percent composition (19.04) but shows maximum positive correlation 0.16 where as 'A' have maximum percent composition (32.72) and shows high negative correlation (-0.21). As 'T' has an average correlation 27.60 and have small negative correlation.

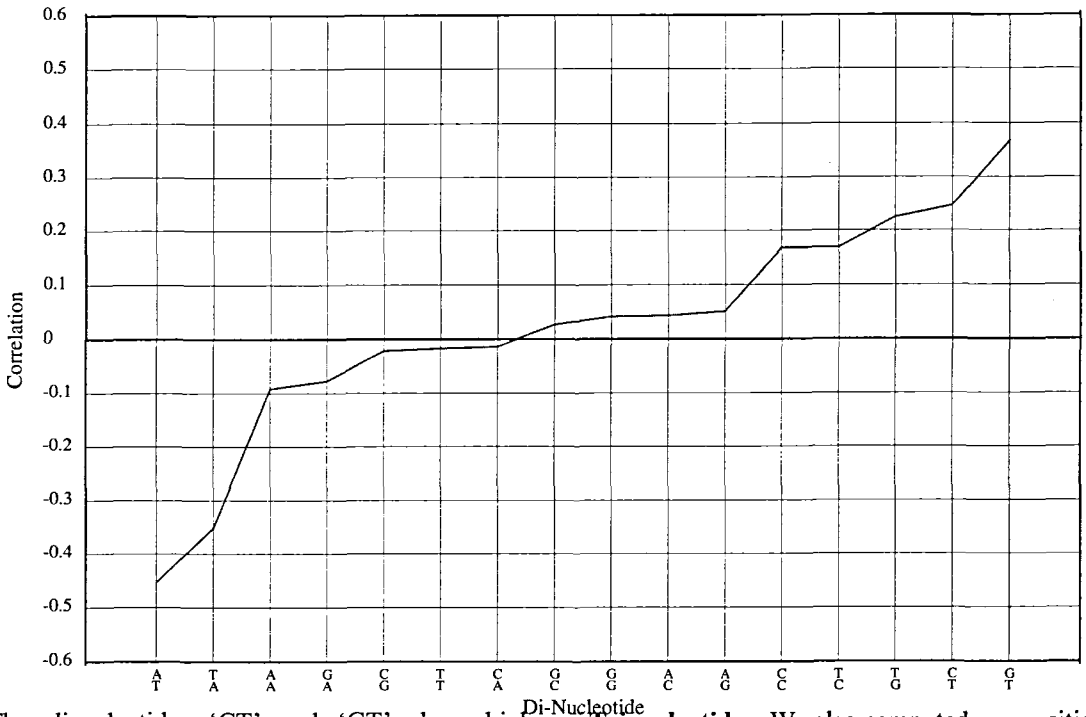
Table 1: Correlation between gene expression and percent composition of nucleotides (single).

Nucleotide	Percent Composition	Pearson Correlation between expression & composition
C	19.04	0.16
G	20.64	0.15
T	27.60	-0.01
A	32.72	-0.21
G+C	39.68	0.23
A+T	61.32	-0.23

In other words, if any nucleotide composition is below than 25% (random) than it have positive correlation and if above 25% than negative correlation. Since yeast genome is GC poor so to make highly expressing genes more distinguishable than other genes, these genes have evolved as GC rich genes. This strategy may be adopted for easy accessibility of protein factors involved in gene expression

Dinucleotide: In the first step, we computed the percent composition of each dinucleotide corresponding to each gene in our dataset using Equation 1 with $s \in \Sigma^2$. Thus we have 16 values (one for each type of dinucleotide) for each gene. In the next step, we calculated the correlation between composition of a dinucleotide and expression level of gene, for each type of dinucleotide. It was observed that some dinucleotides have significant positive correlation, while some others have negative (Figure 1).

Figure 1: Correlation between dinucleotides composition and gene expression level. Y-axis represents correlation between composition of dinucleotides and gene expression and X axis shows the dinucleotides.



The dinucleotides 'CT' and 'GT' show high positive correlation. It is interesting to note that combination of nucleotide 'T' with G and C showing more correlation (maximum) than combination of 'G' and 'C'. Similarly dinucleotide made of A and T gave maximum negative correlation. This means nucleotide 'T' is playing very active role in gene expression, when it made dinucleotide with 'A' then it reduce the gene expression and similarly it increase the gene expression when it combined with 'G' or 'C'. The nucleotide 'T' or dinucleotide 'TT' itself does not show any significant correlation with gene expression level. We also compute correlation between dinucleotides compositions in frame 1 and 2 using Equations 2 and 3, 3 respectively. The correlation trend was same as of overlapping (data not shown).

Trinucleotides: We also computed composition of all possible 64 trinucleotides (codon) in each gene in dataset using Equation 1 with $s \in \Sigma^3$ (overlapping) and then calculate the correlation between composition of codon and gene expression level. As shown in figure 2, there is a large variation in correlation between codon composition and gene expression; a few codons shows high negative correlation (like ATA, GCA, GGA) whereas others shows high positive correlation (like GGT, GCT, GTC). In addition to overlapping codon composition we also computed codon composition in frame 1, 2 & 3 of genes using Equation 4, 5, and 6 respectively and then we compute correlation between codon composition and gene expression. As shown in Figure 2b, 2c and 2d the magnitude of correlation decrease from frame 1 to 3.

Figure 2 (a): The correlation between percent composition of codon and expression level in case of overlapping. The composition was calculated using Equation 1 with $s \in \Sigma^3$.

Figure 2 (b): The correlation between percent composition of codon and expression level in case of frame 1. The composition was calculated using Equation 4.

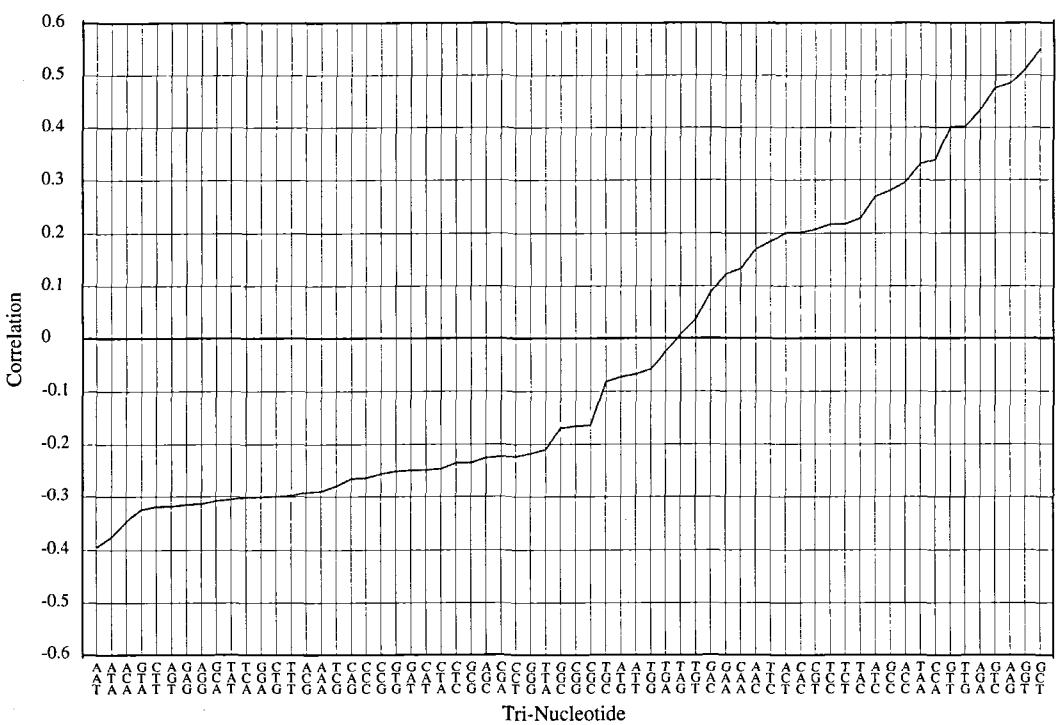


Figure 2 (c): The correlation between percent composition of codon and expression level in case of frame 2. The composition was calculated using Equation 5.

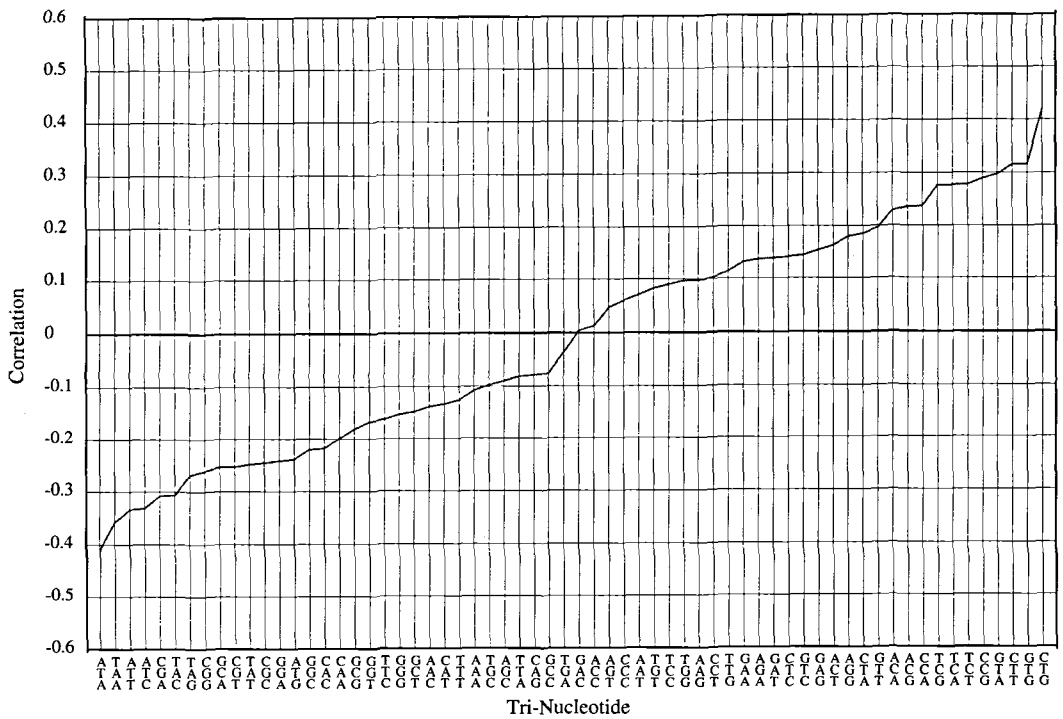
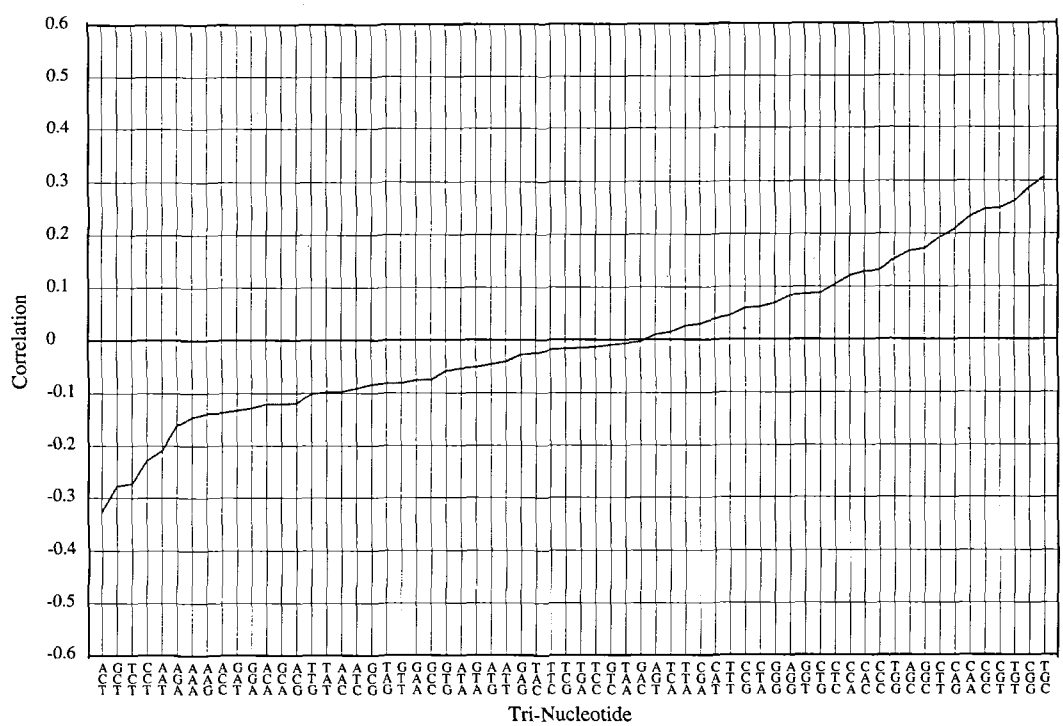
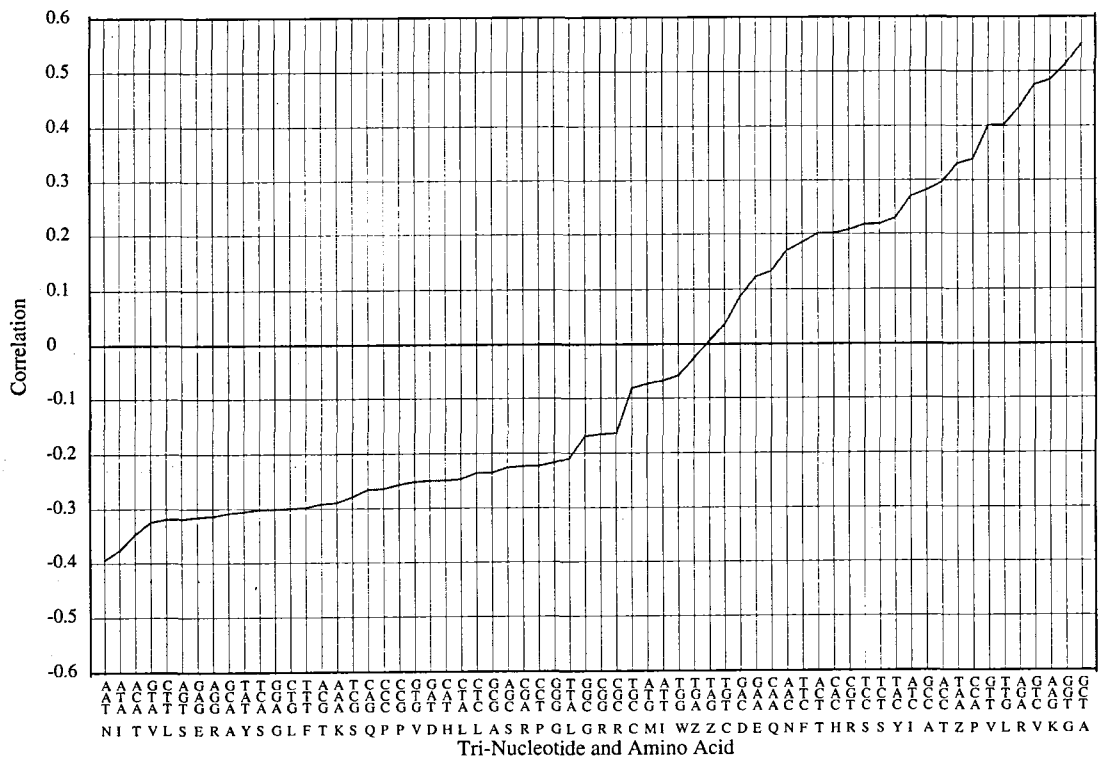


Figure 2 (d): The correlation between percent composition of codon and expression level in case of frame 3. The composition was calculated using Equation 6.



Amino Acid and Codon Composition: It was observed that different codons show different correlation with gene expression. As amino acid uses the synonymous codons so question is whether the different codon used by an amino acid shows same type or correlation or different. Thus we map the amino acid with codon composition. Figure 3 shows the codon correlation with gene expression for first reading frame. As shown in figure different codons of an amino acid have different correlation with expression of gene. That explains the correlation between gene expression and amino acid composition of its protein (Raghava and Han, unpublished).

Figure 3: Correlation between codon composition and gene expression level.



We examined the codon used by amino acids which shows positive correlation with gene expression. Table 2 shows the correlation between gene expression and amino acids obtained from our previous study (http://kiwi.postech.ac.kr/raghava/lgepred/ ; Raghava & Han, unpublished) which also agree with studies of Akashi, 2003 [7]. The table 2 also shows codon in decreasing order of usage of that amino acids having high negative or positive correlation with gene expression.

Table 2. The correlation between percent composition of residues and gene expression level. The residues which have more than +0.2 correlation are shown in shadowed row and residue having correlation (negative) -0.15 are shown by bold letter. The column third shows codons usage of an amino acid (column 1) in *S. cerevisiae* genome (obtained from http://www.kazusa.or.jp/codon/) (22).

Amino Acid	**Pearson Correlation**	**Codon usage in** *Saccharomyces cerevisiae*
A	0.336	GCT(0.38), GCA(0.29), GCC(0.22), GCG(0.11)
C	-0.003	
D	**-0.168**	**GAT(0.65), GAC(0.35)**
E	-0.061	
F	-0.122	
G	0.215	GGT(0.47), GGA(0.22), GGC(0.19)
H	-0.052	
I	-0.136	
K	0.166	
L	**-0.208**	**TTG(0.29),TTA(0.28), CTA(0.14), CTT(0.13), CTG(0.11), CTC(0.06)**
M	-0.087	
N	**-0.210**	**AAT(0.59), AAC(0.41)**
P	-0.064	
Q	-0.052	
R	0.204	AGA(0.48), AGG(0.21),CGT(0.15), CGA(0.07), CGC(0.06), CGG(0.04)
S	**-0.152**	**TCT(0.26), TCA(0.21), TCC(0.16), AGT(0.16), AGC(0.11), TCG(0.11)**
T	0.008	
V	0.269	GTT(0.39), GTA(0.21), GTC(0.21), GTG(0.19)
W	-0.072	
Y	-0.009	

It was observed that, in general, major codons of positively correlated residues (residues having high positive correlation with gene expression) show positive correlation with gene expression. Similarly, it was observed that in general major codons of negatively correlated residues (residues having high negative correlation with gene expression) shows negative correlation. This explains the relation between amino acid and gene expression.

Discussion

In this study we learn the relation between gene expression and nucleotide compositions of gene in a given condition to derive the rules for prediction. We took Hostleage et al. 1998 as reference data because it is well studied and clean (6). We chose *Saccharomyces cerevisiae* because it was widely studied in the past (5). We studied the correlation between nucleotide composition and expression level in detail. As shown in Table 1, this genome is AT rich and shows positive correlations between gene expression and total 'GC' ('G' + 'C') content in a gene. The correlation between gene expression and percent composition is interestingly inversely proportional to percent composition. This means that this organism expressed those genes more which have high amount of nucleotides (like 'G' & 'C') whose overall composition is small in organism. This is interesting, it seems from observation that organism wants to balance the all nucleotides in expressed mRNA. Another interesting feature is nucleotide 'T' which does not show any correlation it self but dinucleotides of 'T' with 'G' or 'C' and with 'A' exhibit high positive and negative correlation respectively. In case of trinucleotides high correlation (positive and negative) was observed for number of codons. This figure agrees with previous observations that expression is biased with percent composition of codon (7-9). As shown in Figure 2 different codon used by amino acids have different correlations with gene expression. This explains the relation between codon biasness and gene expressions. These types of figures provide a better way to understand the relationship between expression and codon usage. It is very interesting that major codon also show positive correlation with gene expression for those amino acids which shows positive correlation with expression level. In contrast, minor codons show positive correlation with gene expression for those amino acids which shows negative correlation with expression. It means minor codons are preferred by some amino acids rather than major codons for high gene expression. This observation is contrast to previous observations where they showed that major codons always have positive correlation with gene expression (7-9). This is due to the definition of codon usage; we used codon usage at genome level(7-9), where as they derived codon usage from highly expressed genes. Thus this contrast is natural because in highly expressed genes these codons (minor codon at genome level which is preferred in expression) will be more than their major codons. As they derived codon usage from these set of highly expressed genes so they consider minor codons (at genome level) as major codons. Our results also explain the evolution where it has been shown in past that few amino

acids is preferred over the other in order to have efficient metabolism (7-9). We observed high correlation between percent composition and expression level for residues Ala, Gly and Val (Table 2) in our previous study(Raghava and Han, unpublished). Interestingly major codons of these amino acids show high correlations with gene expression. This means genes which are responsible to make these proteins having high composition of these amino acids will be preferred in expression. In previous studies it was observed that amino acids made by GNN are preferred(9). Our study also supports this fact and in general this organism supports GC rich genes.

Acknowledgment

The research reported here was supported in part by the Ministry of Information and Communication (MIC) [Foreign Scholar Invitation Program], the Ministry of Science and Technology (MOST) [National R&D Program – Fusion Strategy of Advanced Technologies], and Korea Research Foundation [BK21 Program], of the Republic of Korea.

Reference

1. Zhang, MQ Large-scale gene expression data analysis: a new challenge to computational biologists. Genome Res. 1999 Aug;9(8):681-8.
2. Leung YF, Cavalieri D. Fundamentals of cDNA microarray data analysis. Trends Genet. 2003 Nov;19(11):649-59.
3. Jansen R, Bussemaker HJ, Gerstein M. Revisiting the codon adaptation index from a whole-genome perspective: analyzing the relationship between gene expression and codon occurrence in yeast using a variety of models. Nucleic Acids Res. 2003 Apr 15;31(8):2242-51.
4. Drawid A, Jansen R, Gerstein M. Genome-wide analysis relating expression level with protein subcellular localization. Trends Genet. 2000 Oct;16(10):426-30.
5. Giaever et al Functional profiling of the Saccharomyces cerevisiae genome. Nature. 2002 Jul 25;418(6896):387-91.
6. Holstege FC, Jennings EG, Wyrick JJ, Lee TI, Hengartner CJ, Green MR, Golub TR, Lander ES, Young RA. Dissecting the regulatory circuitry of a eukaryotic genome. Cell. 1998 Nov 25;95(5):717-28.
7. Akashi H. Gene expression and molecular evolution. Curr Opin Genet Dev. 2001 Dec;11(6):660-6.
8. Akashi H, Gojobori T. Metabolic efficiency and amino acid composition in the proteomes of Escherichia coli and Bacillus subtilis. Proc Natl Acad Sci U S A. 2002 Mar 19;99(6):3695-700.
9. Akashi H. Translational selection and yeast proteome evolution. Genetics. 2003 Aug;164(4):1291-303