Available online at www.sciencedirect.com
SCIENCE DIRECT
Biochemical and Biophysical Research Communications 344 (2006) 114–121
BBRC
www.elsevier.com/locate/ybbrc

Relation between mRNA expression and sequence information in *Desulfovibrio vulgaris*: Combinatorial contributions of upstream regulatory motifs and coding sequence features to variations in mRNA abundance

Gang Wu [a], Lei Nie [b], Weiwen Zhang [c,*]

[a] *Department of Biological Sciences, University of Maryland at Baltimore County, 1000 Hilltop Circle, Baltimore, MD 21250, USA*
[b] *Department of Biostatistics, Bioinformatics, and Biomathematics, Georgetown University, Washington, DC 20057, USA*
[c] *Microbiology Department, Pacific Northwest National Laboratory, P.O. Box 999, Mail Stop P7-50, Richland, WA 99352, USA*

Received 14 March 2006

Abstract

The context-dependent expression of genes is the core for biological activities, and significant attention has been given to identification of various factors contributing to gene expression at genomic scale. However, so far this type of analysis has been focused either on relation between mRNA expression and non-coding sequence features such as upstream regulatory motifs or on correlation between mRNA abundance and non-random features in coding sequences (e.g., codon usage and amino acid usage). In this study multiple regression analyses of the mRNA abundance and all sequence information in *Desulfovibrio vulgaris* were performed, with the goal to investigate how much coding and non-coding sequence features contribute to the variations in mRNA expression, and in what manner they act together. Using the AlignACE program, 442 over-represented motifs were identified from the upstream 100 bp region of 293 genes located in the known regulons. Regression of mRNA expression data against the measures of coding and non-coding sequence features indicated that 54.1% of the variations in mRNA abundance can be explained by the presence of upstream motifs, while coding sequences alone contribute to 29.7% of the variations in mRNA abundance. Interestingly, most of contribution from coding sequences is overlapping with that from upstream motifs; thereby a total of 60.3% of the variations in mRNA abundance can be explained when coding and non-coding information was included. This result demonstrates that upstream regulatory motifs and coding sequence information contribute to the overall mRNA expression in a *combinatorial* rather than an *additive* manner.
© 2006 Elsevier Inc. All rights reserved.

Keywords: Sequences; Correlation; Expression

The context-dependent expression of genes is the core for biological activities and large numbers of molecular mechanisms involved in gene regulation have been described during the last decades. With the progress of genome sequencing programs, it is now becoming possible to define various factors within gene regulatory networks and their contributions to gene expression at a global level. Generally, it is believed that gene expression at transcriptional level is mainly controlled by the regulatory motifs present in the promoter regions (sequences upstream of open-reading frame) in a specific spatiotemporal pattern [1,2]. In the past several years, efforts have been made to explore the relationship between gene expression data from microarray analysis and regulatory elements contained in nucleotide sequences identified from genome sequence data [2,3]. These studies provided valuable information for construction of whole genome regulatory networks in various biological systems [2–7]. In particular, a recent study

* Corresponding author. Fax: +1 509 372 1632.
E-mail address: Weiwen.Zhang@pnl.gov (W. Zhang).

0006-291X/$ - see front matter © 2006 Elsevier Inc. All rights reserved.
doi:10.1016/j.bbrc.2006.03.124

showed that it is possible to construct predictive regulatory model based on upstream sequence motif information alone to correctly predict expression patterns for 73% of genes in *Saccharomyces cerevisiae* [8].

In addition to the regulatory motif information contained in the non-coding promoter regions, recent genome-scale analyses suggested a good correlation between gene expression and non-random patterns (such as biased codon usage or amino acid usage) present in coding sequences [9–13]. For example, in yeast, a positive correlation was observed between mRNA expression and codon bias [9] and amino acid usage [10], whereas a negative correlation was found between mRNA abundance and protein length [9]. In *Escherichia coli*, a good correlation between codon usage and gene expression level was observed as well [11,14]. However, since coding sequence (CDS) features and upstream regulatory motifs both partially explain the variations in mRNA abundance, it is unclear whether they explain the same or distinctly different parts of the variations in mRNA abundance. Previously, by using a multiple regression analysis, an attempt was made to predict the mRNA abundance based on a TATA-box score and codon bias in yeast [15]. However, the weighting of the upstream motif contribution with a single TATA-box score may be too simplistic. Furthermore, no attempt was made to determine the relationship between these two types of sequence information in explaining mRNA expression.

Desulfovibrio vulgaris belongs to a group of obligate anaerobic microorganisms, sulfate-reducing bacteria (SRB) [16]. Historical interest in the SRB has been due to their corrosion of pipes and their ability to precipitate heavy metals (e.g., Cr^{6+}, Fe^{3+}, and U^{6+}) and radionuclides (e.g., U^{6+}) from solution via bacterial metal-reduction [17–20]. Recently, the genome of *D. vulgaris* Hildenborough was sequenced [20], which made it possible, for the first time, to investigate gene expression and its regulation in the context of a global cellular network. Our group has been using a whole-genome microarray to study the metabolism of *D. vulgaris* under various growth conditions and its response to environmental stresses, such as temperature variation and oxidative stress [21,22]. Recently a computational analysis of *D. vulgaris* genome was performed and a set of conserved DNA motifs were derived from libraries of potential promoter regions of putative *D. vulgaris* regulons with the AlignACE program [17]. These studies laid a solid foundation for our long-term goals in building a predictive gene expression model in *D. vulgaris*. As one step towards this goal, multiple regression analyses were used here to examine the contribution of regulatory motifs and coding sequence information to the overall gene expression with the objectives to determine: (i) whether information in coding sequences of *D. vulgaris* protein coding genes can also explain some of the variations in mRNA expression; (ii) whether upstream regulatory motifs and coding sequence features contribute to the variation in mRNA expression *additively* or *non-additively*.

Materials and methods

Microarray data. Microarrays containing 3548 ORFs of *D. vulgaris* were designed by NimbleGen System Inc., (Madison, WI) using its Maskless Array Synthesizer (MAS) technology [23]. Arrays were designed with JazzSuite software and the MAS units were used to manufacture the custom arrays. For each ORF, thirteen unique 24-mer oligonucleotides from throughout the ORF were printed onto glass microscope slides. The complete description of experimental design can be found in our previous studies [21]. The raw intensity data were normalized using tools available through the Bioconductor project (http://www.bioconductor.org). The gene calls were based on the Bioconductor implementation of the MAS 5 algorithms [23].

Upstream motif identification and analysis. To retrieve the upstream sequences, complete *D. vulgaris* genome was downloaded from NCBI GenBank (GI: 46562128; Accession No. NC_002937). An ad hoc Perl script was written to extract and format the upstream sequences according to gene attributes (primary annotation) retrieved from TIGR at http://www.tigr.org/tigr-scripts/CMR2/gene-attribute-form.dbi. When the upstream intergenic region was less than 100 bases, all the intergenic sequences were extracted. Otherwise, only the upstream 100 bases were extracted. Upstream regulatory motifs were identified based on the assumption that co-regulated genes (thus in a regulon) are the most likely to share the same regulatory sequence elements [17]. The list of *D. vulgaris* regulons was kindly provided by Prof. Judy Wall and Dr. Chris Hemme of Department of Biochemistry of University of Missouri at Columbia (The regulons were identified based on their homology to the known *E. coli* regulons) [17]. A total of 44 regulons (containing 293 genes) were identified and used in this study. We then used the AlignACE program [8,24] to search over-represented sequence motifs in up to 100 bp upstream of each gene in the same regulon. Motifs with MAP scores greater than 5.0 were regarded as putative conserved motifs [17].

To determine how much of the variations in mRNA abundance can be explained by the upstream motifs, we treated these motifs as independent variables in a multiple regression analysis of mRNA abundance data. For example, if we identify m conserved motifs with the AlignACE program, each gene i ($i = 1 \cdots 293$) will have m independent variable "Motif (i,j)" ($j = 1 \cdots m$). If gene i contains a certain number (n, a positive integer) of motifs j, then Motif $(i,j) = n$. Otherwise, Motif $(i,j) = 0$. Consequently, every gene has a weight for each of the m motifs. Then a multiple regression model was constructed as below:

$$\mathrm{RNA}(i,k,l) = \sum_{j=1}^{m} M_{\mathrm{motif}(j)} * \mathrm{Motif}(i,j) + \mathrm{ID}(k) + \epsilon(i,k,l), \quad (1)$$

where RNA (i,k,l) stands for the microarray intensity (mRNA abundance) of gene i; ID(k) refers to growth condition (LL: lactate log phase, $k = 1$; LS: lactate stationary phase, $k = 2$; FL: formate log phase, $k = 3$; and FS: formate stationary phase, $k = 4$); l (=1,2,3, or 4) refers to replicated measures of mRNA intensity under each growth condition. $M_{\mathrm{motif}(j)}$ is the regression parameter due to the jth motif; ϵ (i,k,l) represents the unexplainable variations. Motifs that significantly contribute to the model fitting were selected through a forward model selection procedure. The significance level for entry into the model is 0.05.

Coding sequences analysis. According to previous studies on the relationship between various CDS features and gene expression [9–13], five types of CDS features (i.e., amino acid usage, codon usage, base composition, gene length, and mRNA secondary structure) were investigated in this study. 3546 coding sequences were downloaded from TIGR as described above. Incomplete ORFs or ORFs containing internal codons (24 CDSs in total) were excluded from the analysis. Consequently, 3522 coding sequences were used in the analysis of CDS features. (1) *Amino acid usage*: correspondence analysis was used to investigate amino acid composition in all 3522 CDSs of *D. vulgaris* with publicly available software *CodonW* (http://codonw.sourceforge.net/), generating the coordinates of each gene on the first four axes (AA-Axis1-4) that summarize the major trends in amino acid usage. To dissect what sequence feature each axis stands for, each axis should be correlated with other *independent* sequence

feature measurements. If there is a strong correlation, it suggests that this axis stands for the sequence feature that the independent measurement measures. In terms of independent measurements of amino acid usage, GRAVY and AROMO were chosen and calculated with *CodonW* [25]. (2) *Codon usage*: relative synonymous codon usage (RSCU) instead of raw counts of codons was used in the correspondence analysis because codon raw counts contain amino acid usage information [26]. Correspondence analysis was performed using *CodonW* and first four axes (CR-Axis1-4) of codon usage were determined. Single-value measures of codon usage used in this study included the effective number of codons (*Nc*) [27] and codon adaptation index (CAI) [14]. In particular, *Nc* was computed with *CodonW*, while CAI was calculated using ribosomal protein genes as the reference set with a web-based application: *The CAI Calculator 2* (http://www.evolvingcode.net/codon/cai/cais.php). (3) *Base composition*: variables measuring base composition in coding sequences include GC (overall frequency of guanine and cytosine), GC3s (frequency of guanine and cytosine at the third codon position for degenerate codons), A3s, T3s, G3s, and C3s (representing the relative frequency of each base at the third codon position). These values were computed with *CodonW*. (4) *Gene length*: the number of bases of each CDS was counted and recorded as gene length. (5) *mRNA secondary structure*: the mRNA secondary structure was predicted with *RNAfold* using the CDS for each gene [28]. The minimum free energy (MFE) of the most possible structure was extracted with an ad hoc Perl script. MFE has also been normalized with gene length as the variable "MFEn" for each CDS.

Two types of statistical analyses, *Pearson's* correlation analysis and multiple regression analysis, were performed for measurements of CDS features and mRNA abundances. *Pearson's* correlation analysis was performed for 3448 genes that have all measures of CDS features and mRNA abundance data (74 CDSs that do not have *Nc* values were excluded) because data distributions in all variables approximately follow normal distributions. We have performed multiple regression analysis for each type of CDS sequence information alone and altogether. Finally, a multiple regression analysis was used to find out whether there were interactions between CDS features and upstream motifs in their contributions to the variations in mRNA abundance. We treated all amino acid usage measures, codon usage measures, base composition measures, gene length, and mRNA secondary structure measures (MFE and MFEn) as five types of independent variables while mRNA abundances under different growth conditions as dependent variables in the multiple regression analysis similar to the model described for motif analysis (Eq. (1)). Multiple regression analysis of mRNA abundance and CDS features was only performed for 293 genes that have been identified in known regulons (see above). All multiple regression analyses were performed with SAS software (SAS Institute Inc., Version 9). All Perl scripts and SAS codes used in this study are available upon request.

Results

The quality of mRNA expression data under different growth conditions

DNA microarray was used in this study to obtain global gene expression data. Four different expression datasets were thus generated corresponding to four growth conditions denoted as LL (lactate-based medium, log phase), LS (lactate-based medium, stationary phase), FL (formate-based medium, log phase), and FS (formate-based medium, stationary phase) [21,29]. Frequency distribution of the mean values of the replicated measurements roughly followed normal distributions after natural logarithm transformation for all four growth conditions [29], indicating that mRNA expression data used in this study are of good quality. *Pearson's* correlation analysis suggested that mRNA expression under FS condition was slightly different from others (Table 1), which may be due to the fact that formate is not the favorable substrate for *D. vulgaris* growth and cell degradation may happen more significantly at stationary phase when grown on this medium [21].

While computational-based motif identification needs information on gene organization and regulation, the information, such as regulons in *D. vulgaris*, is very scarce. In a recent study, a genomic analysis of *D. vulgaris* genome has been done and a short list of putative regulons was generated [17]. In this study, the multiple regression analysis focused on 293 genes that have been determined in these regulons [17] (see below). The frequency distribution of the mRNA abundance of these 293 genes was given in the histogram, which showed a distribution pattern similar to that of the whole genome (data not shown), suggesting that these genes could be considered as a good representative of the whole genome.

Multiple regression analysis indicated that changes in growth conditions contributed to 4.9% of the total variation in mRNA expression of the 293 genes (Table 2, FS 4.6% and LL 0.3%), which suggested that only small amount of change at the mRNA expression level is enough to deal with the changes of energy sources and growth phases in *D. vulgaris*. It is noteworthy that the contribution by growth conditions remained as a constant when other measures of sequence information were combined into

Table 1
Correlation analysis of mRNA abundance of *D. vulgaris* protein coding genes under different growth conditions[*]

	LOG-LL	LOG-FL	LOG-LS	LOG-FS
LOG-LL	1			
LOG-FL	0.825[**]	1		
LOG-LS	0.887[**]	0.801[**]	1	
LOG-FS	0.674[**]	0.731[**]	0.795[**]	1

LOG-LL, LOG-FL, LOG-LS, and LOG-FS are natural logarithm of the mean mRNA abundance measured under LL, FL, LS, and FS conditions, respectively.

[*] Correlation coefficients are calculated based on the data of 3448 protein coding genes.

[**] P value < 0.001.

Table 2
Contribution of sequence features to the variations in mRNA abundance

Sequence features	Percentage of variations explained (%)
Growth conditions	4.9
Upstream motifs alone	54.1
Amino acid usage	23.6
Codon usage	23.9
Base composition	22.0
Gene length	1.3
mRNA secondary structure	6.1
All coding sequence features	29.7
All sequence features	60.3

the multiple regression analysis, suggesting its *independence* of other measures.

Contribution of upstream regulatory motifs to mRNA expression

According to their homologies to members of the 44 regulons defined in *E. coli* [30], 293 *D. vulgaris* genes were identified as belonging to these regulons in a previous study [17]. We used the AlignACE program to search the upstream of these genes up to 100 bp in both strands for putative conserved motifs. As a result, 442 over-represented sequences were identified as conserved motifs. We treated these 442 motifs as independent variables and gave each gene an integer weight according to the times of the presence of a certain motif, therefore generating a motif matrix. Multiple regressions of mRNA abundance data of the 293 genes under four growth conditions against the motif matrix identified 118 motifs that contribute to 54.1% of the variations in mRNA expression (Table 2, and Suppl. Table 1). The top 20 motifs explained around 20% of the total variations in mRNA abundance. Among them, most are involved in important metabolic functions. For example, motif 158 was found upstream of the genes encoding exonuclease ABC (*uvrC* and *uvrB*), DNA ligase (*ligA*), DNA repair protein RecN (*recN*), and RecA protein (*recA*), all of which are involved in DNA metabolism, and motif 338 was found upstream of genes encoding ATP-dependent Clp protease (*clpP*), 60 kDa chaperonin (*groEL*), and heat shock protein HtpG (*htpG*), which are involved in protein fate, and are all highly expressed in *D. vulgaris* [21,22].

Contribution of CDS features to mRNA expression

Amino acid usage

Amino acid composition has previously been shown to correlate with mRNA expression in *E. coli* [25] and yeast [10]. This correlation was interpreted as minimization of metabolic cost during the protein synthesis [31]. Because the amino acids (such as Ala and Gly) with low metabolic cost also have high tRNA concentrations [10], a positive correlation between mRNA abundance and the frequency of these amino acids suggests the presence of selection for translational efficiency [10].

Amino acid usage in the deduced protein sequences of 3522 genes in *D. vulgaris* was analyzed with correspondence analysis to reveal the major trends. The first four trends (AA-Axis1-4) identified by correspondence analysis explain 18.6%, 11.5%, 9.7%, and 7.0% of variability in *D. vulgaris* amino acid usage, respectively. To determine what these major trends stand for, we plotted them against independent measures of sequence patterns such as GC composition in the coding sequence and GRAVY (a measure of hydrophobicity) (Fig. 1) [25]. The first axis (AA-Axis1) correlated with GC composition (Fig. 1A, Suppl. Table 2), suggesting that biased GC content was the major cause of biased amino acid usage. Likewise, strong correlation between the second axis (AA-Axis2) and GRAVY indicated that hydrophobicity was the other predominant determinant of amino acid usage (Fig. 1B, Suppl. Table 2). In other words, the first two major trends in amino acid usage in *D. vulgaris* appear to be the genome base composition and hydrophobicity. Sources for the third and fourth axes (AA-Axis3 and AA-Axis4) appear to be the GC composition as well (Suppl. Table 2).

Consistent with previous observations in *E. coli* and yeast [10,25], we noticed that the AA-Axis1 correlated very well with mRNA abundance in all four growth conditions (Suppl. Table 2), indicative of presence of translational selection in *D. vulgaris* as well. Multiple regression analysis showed that AA-Axis1 alone explained around 20% of the variation in mRNA abundances (Suppl. Table 3). Altogether, various measures of amino acid usage accounted for 23.6% of total variations in mRNA expression (Table 2, Suppl. Table 3), which provides the first evidence of strong association between mRNA expression and CDS features in *D. vulgaris*.

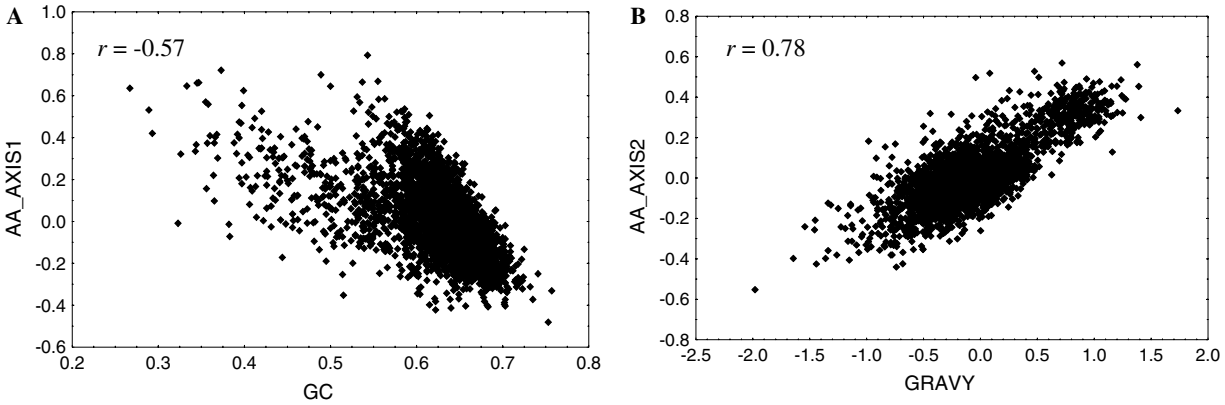

Fig. 1. The first two major trends in amino acid usage of *D. vulgaris* genes represent biased base composition and hydrophobicity, respectively. (A) The coordinate of each gene on the first axis (AA-Axis1) correlates very well with its GC composition. (B) The coordinate of each gene on the second axis (AA-Axis2) strongly correlates with GRAVY, a measure of hydrophobicity.

Codon usage

Early observation of a good correlation between codon usage and gene expression level suggested the action of selection for translational efficiency [14]. Recent large-scale analysis suggested that the relation between codon usage and mRNA expression may differ between genes of different degrees of codon bias [11]. That is, a negative correlation exists between codon bias and gene expression for genes of low degree of codon bias in contrast with a positive correlation for genes of high codon bias. This finding suggests a more complicated relationship between gene expression and CDS information than previously thought. Under this context, we analyzed the codon usage in *D. vulgaris* and its relation to the mRNA expression.

The first four major trends in codon usage (CR-Axis1-4) determined by correspondence analysis accounted for 17.1%, 4.4%, 3.9%, and 3.7% of the total variations in relative synonymous codon usage (RSCU). The first axis (CR-Axis1) strongly correlates with CAI ($r = 0.92$, Fig. 2A). Because CAI has been well documented to correlate well with gene expression level [11–13], the correlation between CR-Axis1 and CAI can be interpreted as evidence of selection for translational efficiency. This interpretation is consistent with the fact that CAI does correlate well with actual mRNA abundance under LL and LS conditions in *D. vulgaris* (Suppl. Table 2). Interestingly, the first axis also strongly correlates with GC3s ($r = 0.94$, Fig. 2B) and GC ($r = 0.73$, Suppl. Table 2), suggesting that biased base composition (due to mutational bias) has also contributed to the codon bias in *D. vulgaris*. Trends on the second, third, and fourth axes (CR-Axis2-4) correlate with the G3s or T3s (Suppl. Table). Therefore, it seems that major trends in codon usage reflect both translational selection and mutational bias in *D. vulgaris*.

Multiple regression analysis indicated that all measures of codon usage were responsible for 23.9% of the variations in the mRNA expression of the 293 genes in known regulons (Table 2, Suppl. Table 3). Interestingly we found in this study that codon usage correlated more strongly with mRNA expression only when growing on the medium with its preferred substrate, lactate, as sole carbon source (CR-Axis1 vs. Log-LL or vs. Log-LS, Suppl. Table 2), which may be due to the fact that cells grown slower on formate- than on lactate-based media because approximately four-fold greater biochemical work required to convert formate into biomass than lactate [32]. The results suggested that a good correlation between codon usage and gene expression may only be observed in fast-growing cells [33].

Base composition

Our analysis showed that major trends in amino acid composition and codon usage strongly associated with the biased base composition (Figs. 1A and 2B). Therefore, it raises the question whether we can simply use base composition as the summary of CDS features. As we found, base composition alone explains 22.0% of the variation in mRNA abundance with the C3s explains the most (13.3%, Suppl. Table 3). However, base composition was not able to represent all CDS features because there are still significant amounts (7.7%) of the mRNA variations that were not represented by the base composition alone (Table 2 and Suppl. Table 3).

Usually, the frequency of G and C was simply combined while we measure the base composition of a genome because of the complementarity of DNA molecules. However, we found that the C3s showed a strong positive correlation with mRNA abundance under all growth conditions, whereas the G3s showed a negative correlation with mRNA abundance (Suppl. Table 2). This observation suggests that it is necessary to examine each base separately when we investigate the relation of base composition and gene expression because only the sense DNA strand is under natural selection due to the uni-directionality of mRNA transcription and protein translation.

Gene length

A negative correlation was found between mRNA abundance and protein length previously in *S. cerevisiae* [9].

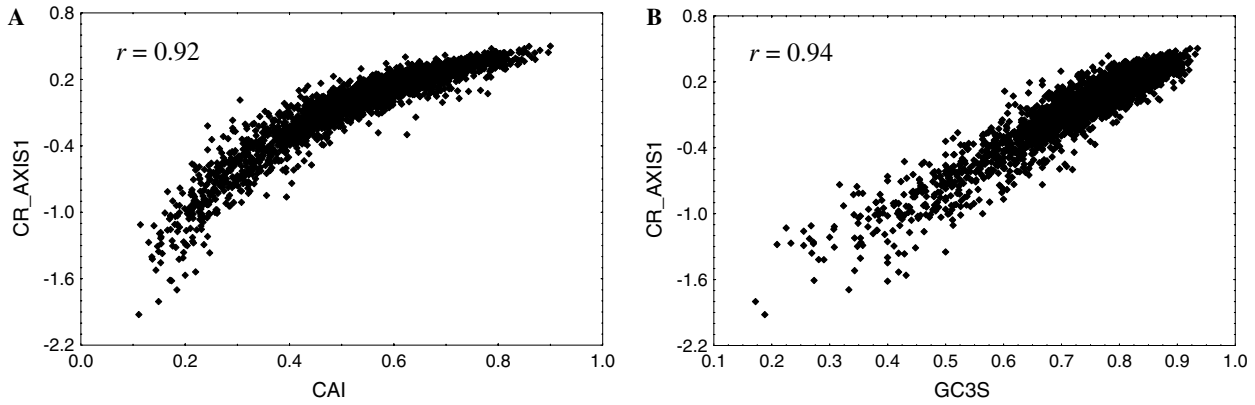

Fig. 2. The major trend in codon usage of *D. vulgaris* genes reflects translational selection and biased base composition. (A) The coordinate of each gene on the first axis (CR-Axis1) correlates strongly with CAI, a codon usage measure that suggests translational selection. (B) CR-Axis1 also strongly correlates with GC composition at the synonymous site (GC3s).

Therefore, we considered gene length as an independent measure of CDS features. Although we did see that the gene length strongly correlated with codon bias (measured with CR-Axis1 and CAI) as previously reported in *E. coli* [34], no significant correlation between gene length and mRNA abundance was found in our study (Suppl. Table 2). Overall, only 1.3% of the variations in mRNA abundance of 293 genes can be attributed to the gene length (Table 2 and Suppl. Table 3).

mRNA secondary structures

It is well known that the mRNA secondary structure can affect mRNA stability and its half life [35,36], but few studies have been performed concerning the relationship between the tendency to form mRNA secondary structures for a given gene and its mRNA abundance. The mRNA secondary structures of all protein coding genes in *D. vulgaris* were predicted with *RNAfold* [28] and the minimum free energy (MFE) of the most possible structure was used as the estimate of the tendency of each gene to form mRNA secondary structure. It appeared the MFE of each gene was highly correlated with the gene length ($r = -0.99$, $p < 0.001$). Therefore, the MFE was normalized to gene length and recorded as MFEn, which was then correlated very well with GC composition ($r = -0.85$, $p < 0.001$, Suppl. Table 2). This is quite normal because the calculation of MFE was based on the interaction between each bases [28]. Multiple regression analysis showed that MFEn explained a small portion (6.1%) of the variations in mRNA abundance of the 293 genes in known regulons (Table 2 and Suppl. Table 3). It is noteworthy that in our current study we only use coding sequences to predict the mRNA secondary structures. However, because of the difficulty to define upstream and downstream mRNA boundaries, our analysis represented so far the most conservative approximation of this characteristic.

All coding sequence features

To determine whether various non-random patterns in coding sequences reflect different sources of variations in mRNA abundance, we performed a multiple regression analysis by including of all variables measuring CDS features. Interestingly, these variables only explained 29.7% of total variation (Table 2 and Suppl. Table 3). This number is far less than the sum of their individual contributions, suggesting that they represented overlapping sources of variations in mRNA abundance.

A combined model which integrates all sequence information

To find out if the variations explained by CDS features are part of those explained by upstream regulatory motifs, we have integrated all the sequence information into the multiple regression analysis. Although upstream regulatory motifs and coding sequences per se explain 54.1% and 29.7% of variations in mRNA abundance, altogether they account for only 60.3% of the total mRNA variations (Table 2 and Suppl. Table 4), suggesting that upstream regulatory motifs and coding sequences explain significant amounts of common variations in mRNA abundance.

Discussion

To dissect the contributions of upstream regulatory motifs and various features in coding sequence to the mRNA expression, we have performed a comprehensive and systematic analysis of their relationship with microarray gene expression data collected from *D. vulgaris*. To assure the quality of this analysis, we first showed that the microarray data were of high quality by demonstrating that the mRNA abundance data under all four growth conditions roughly follow a normal distribution, similar to what has reported in other organisms [11,12]. In addition, we found that the contribution of the growth conditions (in particular FS and LL) remains small but as constants in our multiple regression analyses (Table 2, Suppl. Tables 4 and 5). The constant contribution of growth conditions suggests that the effects of environmental factors on the mRNA expression are *additive* to that of the sequence features.

Upstream regulatory motifs have been successfully applied to explain the variations in microarray data in yeast [37,38] and prokaryotes [39]. In a previous study, the AlignACE program has been successfully used to identify regulatory motifs in GC-rich species *D. vulgaris* [17]. Using the similar approach, the AlignACE program was used to search upstream sequences of 44 regulons (including 293 genes) defined in *D. vulgaris* [17], which allowed identification of 442 over-represented sequence motifs in *D. vulgaris*. Each motif was weighted by the times of its presence upstream of each gene and the weighting value was used as independent variables during the multiple regression analysis of mRNA abundance data. As a result, we found that more than half (54.1%) of the variations in mRNA abundance were able to be explained by the presence of upstream regulatory motifs. This is consistent with previous observation in yeast [8], supporting the concept that the mRNA expression is mainly regulated by the upstream motifs.

In the study, the relationship between mRNA abundance and five types of CDS features was also investigated. It appeared that the CAI was the measure that correlated most strongly with the mRNA abundance ($r = 0.41$, $p < 0.001$) among all single-value measures (i.e., excluding axis values from the correspondence analysis) of CDS features (Suppl. Table 2). This was consistent with previous comparisons of various CDS measures in predicting mRNA expression in yeast [9]. In terms of all measures of CDS features, it is noteworthy that the AA-Axis1 correlated very well with mRNA abundance under all growth conditions (Suppl. Table 2), suggesting that amino acid composition can be most informative indicator of mRNA level. This is confirmed by the multiple regression analysis that AA-Axis1 alone explained around 20% of the variations in mRNA abundance (Suppl. Tables 3–5). Although

coding sequences alone contribute to 29.7% of the variations in mRNA abundance, most of contribution by CDS features is overlapping with that from upstream motifs. Therefore, a total of 60.3% of the variations in mRNA abundance can be explained if both coding and non-coding information was used.

Considering the significant contributions of CDS information to the mRNA abundance defined in this study, an immediate question will be how the mRNA expression was affected by CDS features? One plausible explanation could be from evolutionary biology point of view. First, it is well established that translational selection has driven the usage of amino acids and codons to a non-random pattern in prokaryotic genome, especially in those highly expressed genes [14,40]. In other words, there is a good correlation between amino acid usage or codon usage and protein abundance. Second, since transcription and translation are highly coupled cellular processes [41], the expression of genes and synthesis of proteins at these two stages may have been co-evolved, i.e., more abundant proteins usually need higher concentrations of mRNA transcripts [42]. Consequently, it is possible to observe a correlation between mRNA abundance and biased CDS features in the multiple regression analysis. Accordingly, it is not surprising to see a combinatorial contribution of upstream regulatory motifs and CDS features in our multiple regression analysis. Nevertheless, it should be emphasized that the non-random CDS features are results of mutational bias and translational selection rather than the determinants of mRNA transcription.

Although the CDS features are not the explanatory factors of mRNA expression, the association between them can be very useful in improving gene expression prediction because, (1) they are more easily available and well defined than upstream regulatory motifs; (2) each gene has a numerical measure for each CDS feature, but only has a categorical (or discrete) measure of a motif (in terms of the presence); (3) for genes without identifiable motifs, CDS features remain as the best variables to predict gene expression. In the past years, many statistical models have been proposed to predict the gene expression either based on upstream regulatory motifs [8] or CDS features [15,43,44]. In this study, inclusion of the CDS features in the multiple regression analysis increased coverage of around 6.2% more of the mRNA variations. Although at first glance, 6.2% increase in explanation of the mRNA variations appears to be a small improvement, a careful analysis of the variables suggested that CR-Axis3 and AA-Axis1 are the top two factors that summarized more than 25% of the variations (Suppl. Table 4). To our knowledge, this study presents the first model to combine both regulatory motif information and CDS features in explaining gene expression. Due to the lack of upstream motif information in most of *D. vulgaris* genes (i.e., excluding 293 genes studied here), our current model cannot be used to predict mRNA expression level in the whole *D. vulgaris* genome yet. However, this idea can be applied to this genome when more information on regulons is available or to other genomes where there is better understanding of the regulatory network.

In our current analysis, even by including of environmental factors (i.e., growth conditions), only 66% of the variations in mRNA expression can be explained (Suppl. Table 5). There is still quite a large fraction of the variations in mRNA abundance unexplainable. The quality of current model can potentially be improved from three aspects: (1) if the intensity (strength) in addition to the presence of upstream regulatory motifs can be measured, (2) if some other CDS features such as dinucleotide frequency or codon context can be calculated, and (3) if the mRNA decay rate can be estimated. As an initial step to explore the relationship between mRNA abundance and two major types of sequence information, we have demonstrated in this study that upstream regulatory motifs and CDS features contributed to the variations in mRNA abundance in *D. vulgaris* in a combinatorial manner and using both sequence features may improve the quality of quantitative gene expression prediction in the future.

Acknowledgments

The research described in this paper was conducted under the Laboratory Directed Research and Development (LDRD) Program at the Pacific Northwest National Laboratory, a multi-program national laboratory operated by Battelle for the US Department of Energy under Contract DE-AC056-76RLO1830. G. Wu was supported by the award 0317349 from the NSF DBI program Biological Databases and Informatics (PI: Dr. S.J. Freeland). We are grateful to Drs. Judy Wall and Chris Hemme of Department of Biochemistry of University of Missouri at Columbia for providing regulon information of *D. vulgaris*.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.bbrc.2006.03.124.

References

[1] D. Thieffry, A.M. Huerta, E. Perez-Rueda, J. Collado-Vides, From specific gene regulation to genomic networks: a global analysis of transcriptional regulation in *Escherichia coli*, Bioessays 20 (1998) 433–440.
[2] D. Thieffry, From global expression data to gene networks, Bioessays 21 (1999) 895–899.
[3] B. Futcher, Microarrays and cell cycle transcription in yeast, Curr. Opin. Cell Biol. 12 (2000) 710–715.
[4] M. Caselle, F. Di Cunto, P. Provero, Correlating overrepresented upstream motifs to gene expression: a computational approach to regulatory element discovery in eukaryotes, BMC Bioinformatics 3 (2002) 7.
[5] E.M. Conlon, X.S. Liu, J.D. Lieb, J.S. Liu, Integrating regulatory motif discovery and genome-wide expression analysis, Proc. Natl. Acad. Sci. USA 100 (2003) 3339–3344.

[6] F. Gao, B.C. Foat, H.J. Bussemaker, Defining transcriptional networks through integrative modeling of mRNA expression and transcription factor binding data, BMC Bioinformatics 5 (2004) 31.
[7] F.P. Roth, J.D. Hughes, P.W. Estep, G.M. Church, Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation, Nat. Biotechnol. 16 (1998) 939–945.
[8] M.A. Beer, S. Tavazoie, Predicting gene expression from sequence, Cell 117 (2004) 185–198.
[9] A. Coghlan, K.H. Wolfe, Relationship of codon bias to mRNA concentration and protein length in *Saccharomyces cerevisiae*, Yeast 16 (2000) 1131–1145.
[10] H. Akashi, Translational selection and yeast proteome evolution, Genetics 164 (2003) 1291–1303.
[11] M. dos Reis, L. Wernisch, R. Savva, Unexpected correlations between gene expression and codon usage bias from microarray data for the whole *Escherichia coli* K-12 genome, Nucleic Acids Res. 31 (2003) 6976–6985.
[12] R. Jansen, H.J. Bussemaker, M. Gerstein, 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. 31 (2003) 2242–2251.
[13] R.M. Goetz, A. Fuglsang, Correlation of codon bias measures with mRNA levels: analysis of transcriptome data from *Escherichia coli*, Biochem. Biophys. Res. Commun. 327 (2005) 4–7.
[14] P.M. Sharp, W.H. Li, The codon adaptation index–a measure of directional synonymous codon usage bias, and its potential applications, Nucleic Acids Res. 15 (1987) 1281–1295.
[15] A. Pavesi, Relationships between transcriptional and translational control of gene expression in *Saccharomyces cerevisiae*: a multiple regression analysis, J. Mol. Evol. 48 (1999) 133–141.
[16] G. Voordouw, The genus *Desulfovibrio*: the centennial, Appl. Environ. Microbiol. 61 (1995) 2813–2819.
[17] C.L. Hemme, J.D. Wall, Genomic insights into gene regulation of *Desulfovibrio vulgaris* Hildenborough, OMICS 8 (2004) 43–55.
[18] J.K. King, J.E. Kostka, M.E. Frischer, F.M. Saunders, Sulfate-reducing bacteria methylate mercury at variable rates in pure culture and in marine sediments, Appl. Environ. Microbiol. 66 (2000) 2430–2437.
[19] J.R. Spear, L.A. Figueroa, B.D. Honeyman, Modeling reduction of uranium U(VI) under variable sulfate concentrations by sulfate-reducing bacteria, Appl. Environ. Microbiol. 66 (2000) 3711–3721.
[20] J.F. Heidelberg, R. Seshadri, S.A. Haveman, C.L. Hemme, I.T. Paulsen, J.F. Kolonay, J.A. Eisen, N. Ward, B. Methe, L.M. Brinkac, et al., The genome sequence of the anaerobic, sulfate-reducing bacterium *Desulfovibrio vulgaris* Hildenborough, Nat. Biotechnol. 22 (2004) 554–559.
[21] W. Zhang, D.E. Culley, J.C.M. Scholten, M. Hogan, L. Vitiritti, F.J. Brockman, Global transcriptomic analysis of *Desulfovibrio vulgaris* on different electron donors, Antonie van Leeuwenhoek (In press).
[22] W. Zhang, D.E. Culley, M. Hogan, L. Vitiritti, F.J. Brockman, Oxidative stress and heat-shock responses in *Desulfovibrio vulgaris* by genome-wide transcriptomic analysis, Antonie van Leeuwenhoek (In press).
[23] E.F. Nuwaysir, W. Huang, T.J. Albert, J. Singh, K. Nuwaysir, A. Pitas, T. Richmond, T. Gorski, J.P. Berg, J. Ballin, et al., Gene expression analysis using oligonucleotide arrays produced by maskless photolithography, Genome Res. 12 (2002) 1749–1755.
[24] J.D. Hughes, P.W. Estep, S. Tavazoie, G.M. Church, Computational identification of *cis*-regulatory elements associated with groups of functionally related genes in *Saccharomyces cerevisiae*, J. Mol. Biol. 296 (2000) 1205–1214.
[25] J.R. Lobry, C. Gautier, Hydrophobicity, expressivity and aromaticity are the major trends of amino-acid usage in 999 *Escherichia coli* chromosome-encoded genes, Nucleic Acids Res. 22 (1994) 3174–3180.
[26] G. Perriere, J. Thioulouse, Use and misuse of correspondence analysis in codon usage studies, Nucleic Acids Res. 30 (2002) 4548–4555.
[27] F. Wright, The 'effective number of codons' used in a gene, Gene 87 (1990) 23–29.
[28] I.L. Hofacker, Vienna RNA secondary structure server, Nucleic Acids Res. 31 (2003) 3429–3431.
[29] L. Nie, G. Wu, W. Zhang, Correlation between mRNA and protein abundance in *Desulfovibrio vulgaris*: a multiple regression to identify sources of variations, Biochem. Biophys. Res. Commun. 339 (2006) 603–610.
[30] K. Robison, A.M. McGuire, G.M. Church, A comprehensive library of DNA-binding site matrices for 55 proteins applied to the complete *Escherichia coli* K-12 genome, J. Mol. Biol. 284 (1998) 241–254.
[31] H. Akashi, T. Gojobori, Metabolic efficiency and amino acid composition in the proteomes of *Escherichia coli* and *Bacillus subtilis*, Proc. Natl. Acad. Sci. USA 99 (2002) 3695–3700.
[32] J.J. Heijnen, M.C.M. van Loosdrecht, L. Tijhuis, A black box mathematical model to calculate auto-and heterotrophic biomass yields on Gibbs energy dissipation, Biotech. Bioeng. 40 (1992) 1139–1154.
[33] S. Karlin, J. Mrazek, A. Campbell, D. Kaiser, Characterizations of highly expressed genes of four fast-growing bacteria, J. Bacteriol. 183 (2001) 5025–5040.
[34] A. Eyre-Walker, Synonymous codon bias is related to gene length in *Escherichia coli*: selection for translational accuracy? Mol. Biol. Evol. 13 (1996) 864–872.
[35] U. Lundberg, A. von Gabain, O. Melefors, Cleavages in the 5' region of the ompA and bla mRNA control stability: studies with an *E. coli* mutant altering mRNA stability and a novel endoribonuclease, EMBO J. 9 (1990) 2731–2741.
[36] S. Zhang, M.J. Ruiz-Echevarria, Y. Quan, S.W. Peltz, Identification and characterization of a sequence motif involved in nonsense-mediated mRNA decay, Mol. Cell Biol. 15 (1995) 2231–2244.
[37] D.Y. Chiang, P.O. Brown, M.B. Eisen, Visualizing associations between genome sequences and gene expression data using genome-mean expression profiles, Bioinformatics 17 (2001) S49–S55.
[38] E. Segal, R. Yelensky, D. Koller, Genome-wide discovery of transcriptional modules from DNA sequence and gene expression, Bioinformatics 19 (2003) i273–i282.
[39] A.M. Kierzek, J. Zaim, P. Zielenkiewicz, The effect of transcription and translation initiation frequencies on the stochastic fluctuations in prokaryotic gene expression, J. Biol. Chem. 276 (2001) 8165–8172.
[40] M. Bulmer, The selection-mutation-drift theory of synonymous codon usage, Genetics 129 (1991) 897–907.
[41] J. Gowrishankar, R. Harinarayanan, Why is transcription coupled to translation in bacteria? Mol. Microbiol. 54 (2004) 598–603.
[42] B. Futcher, G.I. Latter, P. Monardo, C.S. McLaughlin, J.I. Garrels, A sampling of the yeast proteome, Mol. Cell Biol. 19 (1999) 7357–7368.
[43] S. Karlin, J. Mrazek, Predicted highly expressed genes of diverse prokaryotic genomes, J. Bacteriol. 182 (2000) 5238–5250.
[44] G. Wu, D.E. Culley, W. Zhang, Predicted highly expressed genes in the genomes of *Streptomyces coelicolor* and *Streptomyces avermitilis* and the implications for their metabolism, Microbiology 151 (2005) 2175–2187.