Journal of
Microbial & Biochemical Technology

Nishio et al., J Microb Biochem Technol 2013, 5:3
DOI: 10.4172/1948-5948.1000101

Research Article Open Access

Metabolic Control of the TCA cycle by the YdcI Transcriptional Regulator in *Escherichia coli*

Yousuke Nishio[1]*, Tomoko Suzuki[2], Kazuhiko Matsui[3] and Yoshihiro Usuda[2]

[1]*Institute for Innovation, Ajinomoto Co, Inc, 1-1 Suzuki-cho, Kawasaki-ku, Kawasaki 210-8681, Japan*
[2]*Research Institute for Bioscience Products & Fine Chemicals,Ajinomoto Co., Inc., 1-1 Suzuki-cho, Kawasaki-ku, Kawasaki 210-8681, Japan*
[3]*Corporate Planning Department, Ajinomoto Co., Inc., 15-1, Kyobashi 1-Chome, Chuo-Ku, Tokyo 104-8315, Japan*

Abstract

Understanding the regulation and control of the expression of genes encoding metabolic enzymes is crucial for production using microbes. To overcome technical difficulties involved in identifying regulatory network systems, we designed a DNA motif finding procedure combining transcriptome and genome sequence data. Here, we used the ArcAB two-component system of *Escherichia coli*, which controls genes involved in the TCA cycle and energy metabolism, as a model to identify DNA motifs involved in gene-expression regulation. DNA-array data were used to extract up-regulated genes from Δ*arcA* and Δ*arcB E. coli* strains, and the upstream sequences were subjected to DNA-motif finding. Sequence similarity and conserved residues identified the known ArcA-binding motif and a novel DNA-motif candidate that was estimated to be related to YdcI, a putative LysR-type transcriptional regulator. A hypothetical YdcI-binding motif was found upstream of the *gltA* gene, suggesting that YdcI might control the carbon flux into the TCA cycle. To verify this, L-glutamic-acid production and citrate synthase activity in the *ydcI* gene-amplified strain were investigated. Our findings suggested that YdcI is a transcription factor that regulates the expression of *gltA* and other genes, and controls the carbon flux into the TCA cycle.

Keywords: L-glutamic acid fermentation; *ydcI* gene; DNA motif analysis; Bioinformatics; *Escherichia coli*

Introduction

Rapid advances in DNA-sequencing technology and bioinformatics have so far provided more than 2,000 microbial genome sequences, huge quantities of omics data, and more than 1,000 biological databases [1,2]. A key challenge of the post omics era is how to acquire novel knowledge based on these data.

From the viewpoint of useful substance production based on fermentation technology, yield improvements are necessary and have been achieved through advances in metabolic engineering technology, including releasing metabolic or genetic regulation, eliminating feedback inhibition, and overcoming rate-limiting reactions [3]. For example, 13C-based metabolic-flux analysis provides information from inside the cell that can be used to identify rate-limiting steps [4]. Further, comparative genomics reveals phylogenetically conserved transcriptional regulation and provides important information about metabolic regulation [5]. However, it cannot provide information related to novel transcriptional regulation, as this requires prior knowledge. Systematic Evolution of Ligands by Exponential Enrichment (SELEX) technology or the combination of chromatin immunoprecipitation with DNA microarrays (ChIP-chip) can show the most likely binding sites of each transcriptional factor, and is useful for regulatory network identification [6,7]. Although these analytical approaches are exhaustive, they are less useful in understanding metabolic regulation during fermentation.

The integration of dynamic gene-expression pattern data and static genome sequences allows common DNA-sequence motifs to be extracted from the upstream regions of commonly regulated genes, in order for activated metabolic regulation to be identified [8-10]. If known DNA-sequence motifs are found, the specific DNA-binding protein might be important in the relevant process. However, several difficulties remain with this approach, including the identification of unknown DNA-sequence motifs or those with no annotations of the gene-regulatory region. Moreover, multiple local-alignment tools such as Multiple EM for Motif Elicitation (MEME) or Gibbs sampler have input-sequence length limitations [11].

To overcome these difficulties, we designed an efficient DNA motif-finding system to identify metabolic and genetic regulators of fermentation phases using both transcriptome and genome sequence data, without the need for regulatory information. To validate our method, we chose the ArcAB two-component regulatory system of *Escherichia coli*, as ArcAB regulates the bacterial transition from aerobic to anaerobic growth and controls more than 100 genes involved in the TCA cycle and energy metabolism [12,13].

DNA-array data for both *arcA* and *arcB* gene-deletion strains were used to classify genes with increased expression levels in Δ*arcA* and Δ*arcB* strains [14]. We explored DNA-sequence motifs in the upstream regions of these genes, based on the assumption that they should have more than 55% identity with high average information content (IC). This was successfully applied for the ArcA-binding DNA-motif search, which we evaluated with a known ArcA-binding DNA-sequence motif using Shannon entropy and semi-global alignment with no penalty for end gaps. Furthermore, we identified a putative regulatory network mediated by YdcI that was annotated as a LysR-type transcriptional regulator with an unknown biological function. As the hypothetical YdcI-binding motif was found upstream of the *gltA* gene, we predicted

***Corresponding author:** Yousuke Nishio, Institute for Innovation, Ajinomoto Co., Inc., 1-1 Suzuki-cho, Kawasaki-ku, Kawasaki 210-8681, Japan, E-mail: yousuke_nishio@ajinomoto.com

Received June 10, 2013; **Accepted** June 27, 2013; **Published** July 01, 2013

Citation: Nishio Y, Suzuki T, Matsui K, Usuda Y (2013) Metabolic Control of the TCA cycle by the YdcI Transcriptional Regulator in *Escherichia coli*. J Microb Biochem Technol 5: 059-067. doi:10.4172/1948-5948.1000101

that YdcI regulated the carbon flux into the TCA cycle; we confirmed this by testing L-glutamic acid production and citrate synthase activity in *ydcI* gene amplification or deletion strains. YdcI was shown to control the carbon flux through the regulation of *gltA* expression.

Materials and Methods

Strains, plasmids and culture conditions

The strains and plasmids used in this study are summarized in Table 1. *E. coli* MG1655 Δ*sucA*Δ*ydcI* was constructed by *ydcI* gene deletion from *E. coli* MG1655 Δ*sucA* [15] according to Datsenko and Wanner [16]. Briefly, the PCR primers 5′-aaggagggatcgacagatcccttcacctttcagaacggcattgattttcgtgaagcctgctttttttat-3′ and 5- ggagtgtaggtaaccgcattcactcttgcgggaagaatttacaaactgtgcgctcaagttagtataaa-3′ were used to amplify a fragment that replaced the *cat* gene. The pMW118-λattL-Cm^R -λattR plasmid was used as a PCR template [17] and the pKD46 plasmid was used for Red recombination [16]. Plasmid pMW- λint-xis encoding λXis/Int recombinase was used to eliminate all DNA fragments that were flanked by λattL/R sites [18]. The pMW218-*ydcI* plasmid was then constructed using the PCR fragment containing the *ydcI* gene and its upstream region. The PCR primers 5′-gcggatcctcgaatgtaccggca-3′ and 5′-gcaagcttggagggatcgacaga-3′ were used to amplify a fragment that was digested with *Bam*HI and *Hin*dIII, and ligated into the pMW218 vector.

For manipulation, cells were grown in L-broth containing 10 g BactoTryptone (Difco, Japan), 5 g yeast extract (Difco), and 10 g NaCl/L distilled water, or in T-broth containing 12 g BactoTryptone, 24 g yeast extract, and 8 ml glycerol/L distilled water, adjusted to pH 7.0 with potassium phosphate buffer. Cultivation for L-glutamic acid fermentation was performed for 24 h at 37°C in a 500-ml shaking flask with a working volume of 20 ml in MS medium containing 40 g glucose, 1 g $MgSO_4 \cdot 7H_2O$, 24 g $(NH_4)_2SO_4$, 1 g KH_2PO_4, 10 mg $FeSO_4 \cdot 7H_2O$, 8.2 mg $MnSO_4 \cdot 4H_2O$, and 2 g yeast extract/L distilled water. The MS medium pH was adjusted to 7.0 using KOH. After sterilization, 30 g of $CaCO_3$ was added and growth was monitored by measuring the optical density at 600 nm. Chloramphenicol (30 mg/l) or kanamycin (50 mg/l) was added to media to select for the corresponding markers in the bacterial chromosome or plasmid.L-glutamic acid and glucose concentrations were measured using the biotech analyzer BF-5 (Oji Scientific Instruments, Japan).

Enzyme assay

Citrate synthase activity was measured according to the method of Weitzman [19]. Cells in mid-log phase grown in T-broth were harvested and disrupted with a Multi-beads shocker® (cell disruptor) (Yasui Kikai, Japan) operating at 2,500 rpm, following three cycles of 60 s on and 60 s off. After centrifugationat 15,000×g for 10 min to remove beads, 20 µl supernatant was used as a crude extract in an assay at 37°C. The assay solution also contained 0.1 M Tris-HCl (pH 8.0), 8 mM Acetyl-CoA (Sigma), 10 mM5,5-dithiobis (2-nitrobenzoic acid) (Sigma) and 10 mM oxaloacetate (substrate; Sigma). Activity was monitored by measuring the absorbance at 412 nm on a 96-well plate. The protein concentration of the crude extract was measured using the Bradford method.

Strain	Description	Source or reference
MG1655	Wild type *E. coli* K12	Laboratory stock
Δ*sucA*	MG1655Δ*sucA*::cm	[15]
Δ*sucA*Δ*ydcI*	MG1655Δ*sucA*Δ*ydcI*::cm	This study
Δ*sucA* / pMW218	MG1655Δ*sucA*::cm / pMW218	This study
Δ*sucA* / pMW-*ydcI*	MG1655Δ*sucA*::cm / pMW218-ydcI	This study
Plasmid		
pMW218	vector control for plasmid pMW-*ydcI*	Nippon gene Co., Ltd.
pMW-*ydcI*	pMW218 with *ydcI*	This study

Table 1: Strains and plasmids.

DNA-array data analysis

DNA-array data were obtained from GenoBase [14]. Gene Cluster 3.0 and Microsoft Excel were used for the analyses [20]. To extract candidate genes with increased expression levels in both Δ*arcA* and Δ*arcB* strains, the data were compared with a wild-type strain and genes showing significant differences in expression levels were extracted. Numerical values for the expression rate of each spot were transformed into logarithmic values, and used to calculate the means and standard deviations. Based on the variance-analysis method, we calculated the unbiased variance for each gene, which was used in a *t*-test. The DNA array data had an insufficient sample size, which was overcome by multiplying the *p*-value obtained from the *t*-test by 4. Based on the corrected *p*-value, we evaluated the result of the t-test with the Bonferroni correction as follows:

Unbiased variance (Ve)=Sum of squared deviations (Se) / Degree of freedom (Ne).

Here, Se=sigma (measured value – average)^2, Ne=total spot number – number of experimental condition, Statistic=|D sample mean|/√{Ve× (1/spot number) ×2}, and level of significance=0.05/3.

In the case of wild-type comparison DNA-array data, we excluded genes that showed large differences with multiplex spots within the same DNA array. For other DNA-array data, we selected genes with expression levels that showed that the logarithmic value of the ratio of two spots was not equal to 0. Specifically, we calculated the average and standard deviation of the logarithmic value of each spot, and extracted those genes in Δ*arcA* or Δ*arcB* strains with expression levels greater than average plus two times the standard deviation value. We classified these genes with K-means clustering, and selected those that showed increased expression in both Δ*arcA* and Δ*arcB* strains.

Bioinformatics

C shell and MATLAB (matrix laboratory) were used to make scripts, MEME was used for DNA-motif construction [21], and WebLogo or MATLAB was used for visualization of the DNA motif [22]. Regulatory factor-binding sequences in *E. coli* were obtained from RegulonDB, DPinteract, *Escherichia coli* Transcription Factor Binding Sites (ECTFBS), and EcoCyc [23-26].

Local multiple alignment

An outline of the process is shown in Figure 1. Briefly, 20-mer, 25-mer, and 30-mer lengths of unique sequences were extracted from the input upstream regions using EMBOSS: word count, and identical or similar sequences were identified using EMBOSS: fuzznuc. Similarity criteria were as follows: more than 56% identity against the 30-mer input sequence; more than 60% identity against the 25-mer input sequence; and more than 65% identity against the 20-mer input sequence. Similar sequences were aligned using MEME and treated as a DNA motif and an N-line M-raw character matrix. The conservation quality of each row was evaluated using IC which was calculated using the Shannon

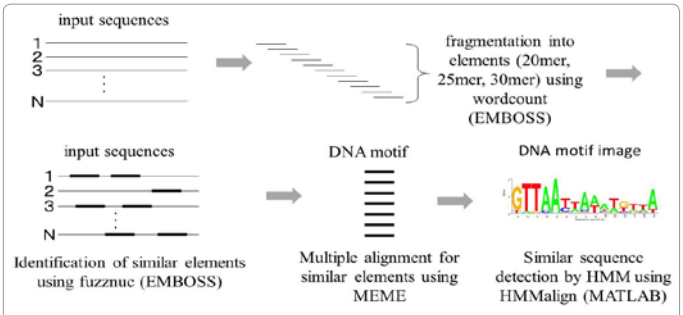

Figure 1: Schematic of locally conserved sequence identification. Input sequences were broken down into 20-mer, 25-mer, or 30-mer length elements using the wordcount program. Sequences of high similarity to each element were identified using fuzznuc and aligned using MEME.

entropy [27]. The IC was calculated by the following equation:

$$\mathrm{IC} = \sum_{b=A}^{T} f_b \log_2 \frac{f_b}{p_b}$$

Here, f_b is the appearance frequency of A, G, C and T in a residue position of the DNA motif, and P_b is the A, G, C, and T frequencies in the *E. coli* genome, which were each set to 25%.Summation of the IC in each row was treated as the information amount of the DNA motif. The conservation quality of each line was evaluated using the identity obtained from pair wise alignment using the Smith-Waterman method [28]. The probability of DNA-motif occurrence was calculated by the binomial distribution as follows:

$$p = \sum_{m=A}^{C} nCm \times 0.25^m \times (1-0.25)^{n-m}$$

Here, p is the probability, n is the number of DNA sequences constituting the DNA motif, and m is the number of A, T, G, or C bases in the DNA motif. The DNA-motif occurrence score was obtained by the logarithm of probability which was multiplied by –1 as follows:

$$Score = -\log p$$

To test whether the DNA motif was constructed from homologous sequences, we performed a statistical value comparison between the DNA motif and a pseudo motif of the same size with a sequence collected randomly from the *E. coli* genome. This showed significant differences of conservation quality of both lines and rows. Toassess sequence similarity, the Wilcoxon signed-rank test was used with a significance level of 0.05. This procedure was repeated 200 times to prevent comparison with a highly conserved pseudo-DNA motif with sequences collected by chance. In more than 150 cases, the DNA motif was shown to be significantly conserved. For conservation quality of rows, on average, an IC>1.0 was used for further analysis of significantly conserved DNA motifs. In a separate statistical test, 200 DNA sequences ranging from 10-mers to 40-mers were picked at random, and their identities calculated. The averaged identity was treated as the true value in the *E. coli* genome. Statistical evaluation regarding the identity between the DNA motif and the true value in the *E. coli* genome was tested using the one variable *t*-test with a significance level of 10^{-8}.

Clustering analysis of DNA motifs

Similarity scores between DNA motifs were calculated using the IC of the DNA motif and semi-global alignment with no penalty for end gaps (Figure 2). Scores were calculated by the recurrence formula:

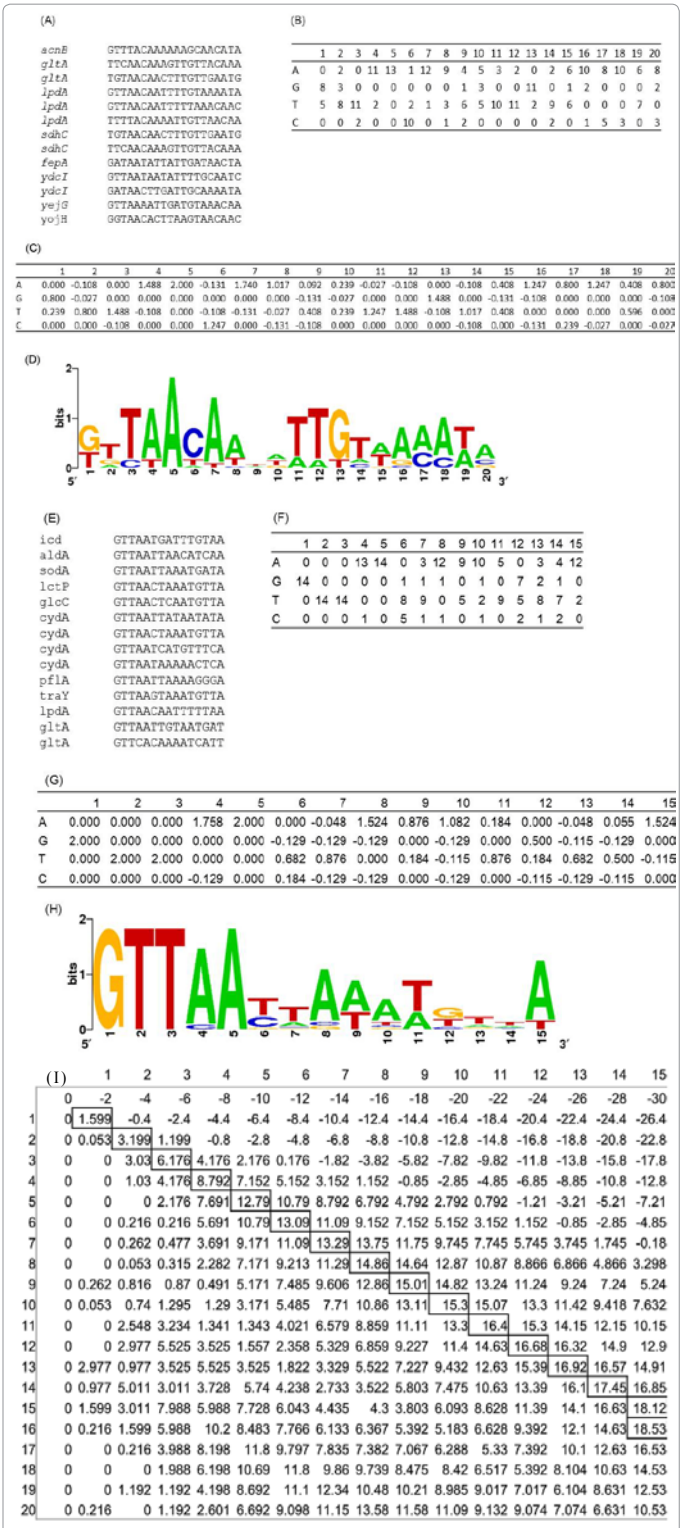

Figure 2: DNA-motif comparison.(A) Extracted 20-95-3 DNA motif. (B) Frequencies of A, G, T and C in each residue of DNA motif (20-95-3). (C) IC matrix of DNA motif (20-95-3) obtained by Shannon entropy. (D) Graphical representation of DNA motif (20-95-3). (E) Known ArcA-binding motif (ArcA-X; 29). (F) Frequencies of A, G, T and C in each residue of ArcA-binding motif. (G) IC matrix of ArcA-binding motif obtained by Shannon entropy. (H) Graphical representation of ArcA-binding motif. (I) Scoring matrix for obtaining maximum homology between DNA motifs using semi-global alignment with no penalty for end gaps.

$$\text{score}(i,j)=\max\begin{cases} score(i-1,j-1)+\sum_{b=A}^{T} abs\left(IC(i)_b^{motif\,1}\times IC(j)_b^{motif\,2}\right) \\ score(i-1,j)-d \\ score(i,j-1)-d \end{cases}$$

Here, $IC\ (i)_b^{motif1}$ is the IC of A, G, C or T bases on the i residue of the first DNA motif, $IC\ (j)_b^{motif2}$ is the IC of A, G, C or T bases on the j residue of the second DNA motif, and *d* is the gap penalty which is set to 2.The distance matrix of the DNA motifs was calculated as the inverse of the similarity matrix. Clustering using the distance matrix was performed with SOM Toolbox for MATLAB (http://www.cis.hut.fi/projects/somtoolbox/).

Identifying genome sequences similar to the DNA motif

A hidden Markov model (HMM) of the DNA motif was constructed using the hmmprofstruct and hmmprofestimate functions in MATLAB; the hmmprofalign function was used to find genome sequences similar to the DNA motif. To confirm that the detected sequence was significantly similar to the input DNA motif, we compared the similarity between the detected and random sequences. To obtain the random-sequence similarity score, the *E. coli* genome was split into 100-bp fragments, and the most similar sequences were selected from each spliced region. These were treated as random sequences, and their averages and standard deviations were calculated using the alignment score of the hmmprofalign function. We used the z-test to determine whether the score of the homologous sequence was higher than that of the random sequence.

Results

Analysis of DNA-array data

We analyzed the DNA-array data of the Δ*arcA* or Δ*arcB* strains in GenoBase [14]. This identified 101 genes with expression levels that differed significantly between a wild-type and Δ*arcA* strain or between a wild-type and Δ*arcB* strain. We classified these gene-expression profiles into seven clusters using the K-means clustering method (Figure 3). Twenty-one genes from one of the seven clusters showed increased expression in both Δ*arcA*-and *arcB*-deletion mutants (Figure 3A), suggesting that their expression was directly or indirectly controlled by the ArcAB system. Of these 21 genes, 17 (*acnB*, *acs*, *bglJ*, *fepA*, *glcB*,*gltA*, *hdeA*, *icd*, *lpdA*, *mqo* (formaly*yojH*), *osmB*, *purE*, *sdhC*, *ydcI*,*yejG*, *yhhX*, and *ycgF*) were the first genes of an operon or had an independent transcription unit, and ArcA-binding sites were expected to be located in their upstream regions. ArcA-binding sites have previously been shown to exist upstream of *acnB, glcB, gltA, icd, lpdA, sdhC*, and *sucC* [23], but the other genes are not known members of the ArcA regulon. We hypothesized that these were either unidentified members of the ArcA regulon or were not directly regulated by the ArcAB system. To identify ArcA-binding sites of the 17 genes that lacked regulatory sequence annotations, we analyzed 500-bp upstream and 100-bp downstream of the start codon (Figure 1). We analyzed *glcD* instead of *glcB*, as the former was the first gene of the *glc* operon.

DNA-motif discovery

We attempted to extract all conserved DNA motifs and compare them with known ArcA-binding motifs. We assumed that more than one-half of the input upstream regions of the 17 genes would possess an ArcA-binding site, and conditioned more than nine similar sequences to be extracted as such. The fuzznuc program showed that 68 of 10,398 elements for 20-mer sequences and 101 of the 10,218 elements for 30-mer sequences possessed more than nine similar sequences. The 68 and 101 sets were aligned using MEME, and 20 of the 507 DNA motifs were extracted as statistically significant and classified using SOM clustering (Figure 4). As a positive control, we included the ArcA-binding motif proposed by McGuire et al. [29] (ArcA-X in Figure 4), and the modified ArcA-binding motif that removed non-*E. coli* sequences from McGuire's ArcA-binding motif (ArcA-Y in Figure 4). As a negative control, we included the Mlc-binding motif, the sequences of which were obtained from the EcoCyc database [23].

No extracted DNA motif was classified into exactly the same cluster as the positive controls; however, DNA motif 20-95-3 was classified into a nearby cluster. We used semi-global alignment with no penalty for end gaps to compare the ArcA-binding motif with 20-95-3 using the DNA motif IC (Figure 2). Six of the 20 DNA motifs (20-64-3, 20-379-3, 20-520-1, 20-520-3, 20-305-1, and 20-8-1) were classified into a distant cluster to the ArcA-binding motif, and their sequences were completely different from those of the 14 remaining DNA motifs

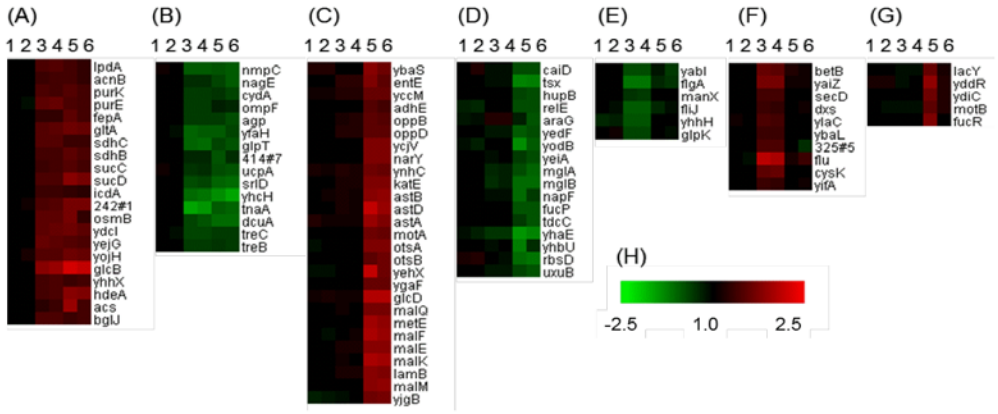

Figure 3: K-means clustering analysis of DNA-array data.Original data from GenoBase (14). Two spots of the same gene were compared on a slide: 1 and 2, wild-type spots; 3 and 4, wild-type and Δ*arcA* strains; 5 and 6, wild-type and Δ*arcB* strains. Red denotes messenger RNA (mRNA) increase in gene-deletion strain. Green denotes mRNA decrease in gene-deletion strain. (A) Genes with increased expression levels in both Δ*arcA* and Δ*arcB* strains. (B) Genes with decreased expression levels in both Δ*arcA* and Δ*arcB* strains. (C) Genes with increased expression levels in Δ*arcB* strain. (D) Genes with decreased expression levels in Δ*arcB* strain. (E) Genes with decreased expression levels in Δ*arcA* strain. (F) Genes with increased expression levels in Δ*arcA* strain. (G) Genes with increased expression levels in Δ*arcB* strain at only one spot. (H) Color bar indicating expression level.

(Figure 4). The ArcA-binding motif and 20-95-3 had an alignment score of 18.53, revealing their similarity with each other (Figure 2).

Sequence composition of DNA motif and comparison with known ArcA-binding sequence

As shown in Table 2, the 20-mer 20-95-3 DNA motif was identified upstream of the following eight genes: *acnB*, *fepA*, *gltA*, *lpdA*, *mqo*, *sdhC*, *ydcI*and *yejG*. Of the 17 genes identified by DNA-array analysis, *acnB*, *gltA*, *lpdA*, *sdhC*, *glcB* (*glcD*), and *icd* are known ArcA-regulated genes (or operons). The known ArcA-binding sequences in the upstream region of *acnB*, *gltA*, *lpdA*, and *sdhC* partially corresponded to the sequence of the 20-95-3 DNA motif (Table 2), indicating that the motif contained several ArcA-binding sequences but that not all known binding sites could be detected. Therefore, we constructed an HMM of the 20-95-3 DNA motif to identify similar sequences. As the IC of the 20-95-3 DNA motif ninth and tenth residues was almost 0, we constructed two types of HMM with profile lengths of 20 and 18, and examined three *p*-values (0.01, 0.001, and 0.0001). Table 3 shows that similar sequences to the 20-95-3 DNA motif were identified in the upstream regions of *acs*, *osmB*, *glcB* (*glcD*), and *ycgF*, and that we found no putative ArcA-binding sequence in the upstream region of *icd*, *yhhX*, *hdeA*, *purE*, or *bglJ*, although the detection failure in the case of *icd* might have been due to analytical error, as an ArcA-binding site in the upstream region of *icd* has previously been experimentally validated [30].

DNA-motif identification excluding the ArcA-binding motif

We explored the possibility of the existence of a different gene network from the ArcA regulon. We noted that the expression level of *ydcI* was significantly increased in both Δ*arcA* and Δ*arcB* strains, and that two ArcA-binding sequences were predicted upstream of *ydcI* (Table 2). YdcI was annotated as anLysR-type DNA-binding protein with unknown biological function. We hypothesized that YdcI also controlled the expression of several genes. We searched for different DNA motifs from 20-95-3 using the method used for the discovery of the ArcA-binding motif. As more than nine sequences had already been analyzed, we selected DNA motifs containing four to eight sequences. Eleven of the 507 DNA motifs were extracted as being statistically significant, and were classified using SOM clustering (Figure 5). As before, ArcA-X and ArcA-Y (Figure 5) and the Mlc-binding motif were included as negative controls for SOM clustering [23,29]. Ten of the 11 DNA motifs were not classified into the same cluster as the ArcA-binding motifs or the Mlc-binding motif, suggesting that they lacked similarity. Of the 10 motifs, we focused on 20-150-1, which was located upstream of *hdeA*, *gltA*, *icd*, *lpdA*, *sdhC*, and *yhhX* (Table 4). The upstream region of *hdeA*, *icd,* and *yhhX* did not possess the 20-95-3 sequence or its homolog sequences. We found similar sequences to 20-150-1 in the upstream regions of *acnB*, *fepA*, *glcD* (*glcB*), *gltA*, *hdeA*, *icd*¸ *lpdA*, *purE*, *sdhC*, and *yhhX* by constructing an HMM of 20-150-1 and using three *p*-values (0.01, 0.001, and 0.0001) (Table 5). Based on this result, we proposed a gene-regulatory network induced by YdcI (Figure 6).

We treated the nucleotide sequences of the DNA motif 20-150-1 as the predicted YdcI-binding sequence, and annotated the upstream

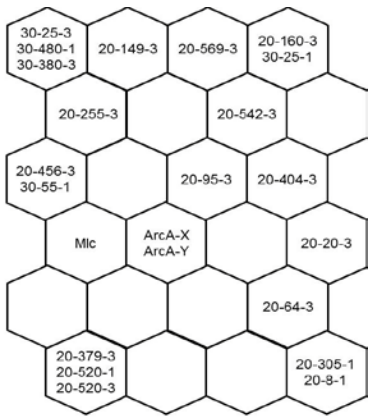

Figure 4: DNA-motif classification by SOM clustering for ArcA-binding motif. SOM clustering result was shown in hexagonal lattice. Similar DNA motifs are in the same hexagon. ArcA binding motif and Mlc binding motif were used as positive and negative control, respectively.

Sequence (5′-)		
acnB	GTTTACAAAAAAGCAACATA	known ArcA-binding sequence
gltA	TTCAACAAAGTTGTTACAAA	known ArcA-binding sequence
gltA	TGTAACAACTTTGTTGAATG	known ArcA-binding sequence
lpdA	GTTAACAATTTTGTAAAATA	known ArcA-binding sequence
lpdA	GTTAACAATTTTTAAACAAC	known ArcA-binding sequence
lpdA	TTTTACAAAATTGTTAACAA	known ArcA-binding sequence
mqo	GGTAACACTTAAGTAACAAC	unknown sequence[a]
sdhC	TGTAACAACTTTGTTGAATG	known ArcA-binding sequence
sdhC	TTCAACAAAGTTGTTACAAA	known ArcA-binding sequence
fepA	GATAATATTATTGATAACTA	unknown sequence[a]
ydcI	GTTAATAATATTTTGCAATC	unknown sequence[a]
ydcI	GATAACTTGATTGCAAAATA	unknown sequence[a]
yejG	GTTAAAATTGATGTAAACAA	unknown sequence[a]

[a]*fepA*, *mqo*, *ydcI*, and*yejG* genes are not known members of the ArcAregulon.

Table 2: DNA motif 20-95-3 sequences and their verifications.

Gene	HMM motif length					
	20 bp			18 bp		
	p-value			*p*-value		
	0.0001	0.001	0.01	0.0001	0.001	0.01
acnB	Y	Y	Y	N	N	N
acs	N	N	N	N	N	Y
bglJ	N	N	N	N	N	N
fepA	Y	Y	Y	N	N	N
glcD	N	Y	Y	N	N	N
gltA	Y	Y	Y	Y	Y	Y
hdeA	N	N	N	N	N	N
icd	N	N	N	N	N	N
lpdA	Y	Y	Y	Y	Y	Y
mqo	Y	Y	Y	N	N	N
osmB	N	N	N	N	Y	Y
purE	N	N	N	N	N	N
sdhC	Y	Y	Y	Y	Y	Y
ycgF	N	N	Y	N	N	N
ydcI	N	N	N	N	N	N
yejG	Y	Y	Y	N	N	N
yhhX	N	N	N	N	N	N

Y, homologous sequence found; N, homologous sequence not found

Table 3: Identification of homologous sequences to DNA motif 20-95-3 using HMM.

regions of 10 genes using EcoCyc [23] (Figure 7). We found that the predicted YdcI-binding site partially overlapped the P1 promoter in the upstream region of *gltA* (Figure 7A), was located adjacent to the P2 promoter upstream of *icd* (Figure 7B), and was located just upstream of the ArcA-binding site upstream of *sdhC* (Figure 7C). These results suggested that YdcI is involved in the regulation of expression of *gltA*, *icd*, and the *sdh* operon.

Experimental validation of the gene-regulatory network prediction

The flavor enhancer L-glutamic acid is produced worldwide in large quantities, and is an important fermentation product. Previously, we constructed a simulation model of the *E. coli* L-glutamic acid-producing strainMG1655 Δ*sucA*, and found that the expression of *gltA* and *icd* was highly sensitive to L-glutamic acid production [31]. We hypothesized that if YdcIactivates *gltA* or *icd* gene expression, then *ydcI* gene amplification should lead to an increase in L-glutamic

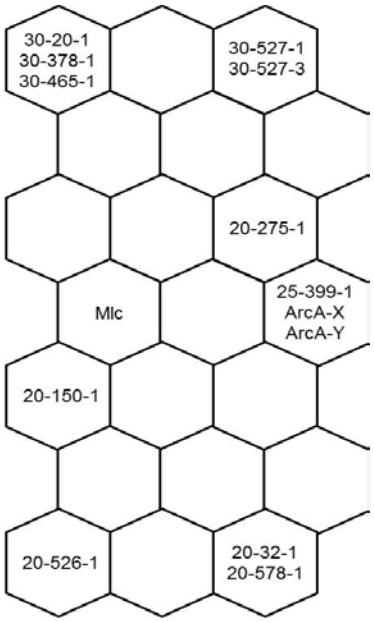

Figure 5: DNA-motif classification by SOM clustering for YdcI-binding motif. SOM clustering result was shown in hexagonal lattice. Similar DNA motifs are in the same hexagon. ArcA binding motif and Mlc binding motif were used as positive and negative control, respectively.

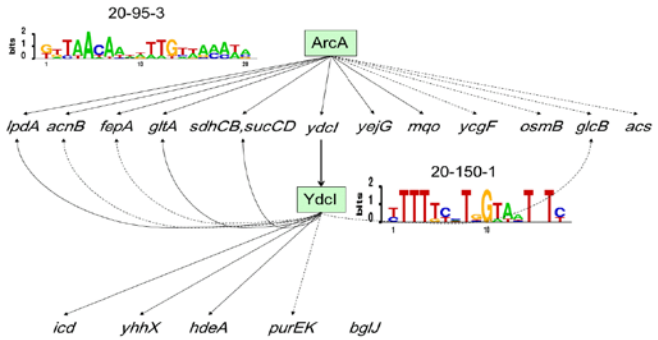

Figure 6: Proposed gene-regulatory network.Solid arrow indicates that the composed sequence of the DNA motif was presented upstream of the genes. Dashed arrow indicates that similar sequences of DNA motifs were presented upstream of the genes. Bold arrow shows *ydcI* gene transcription to YdcI protein.

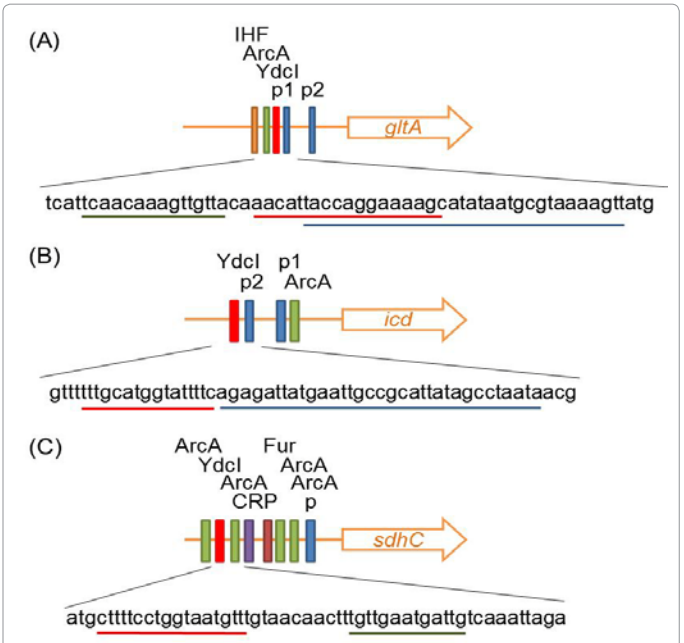

Figure 7: Predicted YdcI-binding sites upstream of *gltA*,*icd*, and *sdhC*. Annotation of known regulatory sequences was derived from EcoCyc. (A) The predicted YdcI-binding site overlapped annotation of the regulatory region for *gltA*. The ArcA-binding site sequence is underlined in green (tcaacaaagttgtta), the predicted YdcI-binding site in red (aacattaccaggaaaag), and the promoter sequence in blue (taccaggaaaagcatataatgcgtaaaagt). (B) The predicted YdcI-binding site overlapped annotation of the regulatory region for *icd*. The sequence of the predicted YdcI-binding site is underlined in red (ttttgcatggtattttc), and that of the promoter sequence in blue (agagattatgaattgccgcattatagcctaata). (C) The predicted YdcI-binding site overlapped annotation of the regulatory region for*sdhC*. The sequence of the ArcA-binding site is underlined in green (ttgttgaatgattg), and that of the predicted YdcI binding site in red (cttttcctggtaatgtt).

Gene	Sequence
gltA	cttttcctggtaatgtt
hdeA	tttttcatcgtaatatc
icd	ttttgcatggtattttc
lpdA	ttttttctggtaatctc
sdhC	cttttcctggtaatgtt
yhhX	tttttttttgatctttc

Table 4: Nucleotide sequence of DNA motif 20-150-1.

acid production, and if YdcI represses *gltA* or *icd* gene expression, then L-glutamic acid production should decrease. To test this experimentally, we constructed a*ydcI* gene-amplification strain using a low-copy-number plasmid (Δ*sucA*/pMW-*ydcI*) and the *ydcI* gene-deleted strain (Δ*sucA*Δ*ydcI*) from *E. coli* MG1655 Δ*sucA*. L-glutamic acid production was evaluated by shaking-flask cultivation (Table 6).

Cellular growth of all strains was similar, and residual sugar levels were close to 0. L-glutamic acid accumulation of the *ydcI* gene-amplified strain (Δ*sucA*/pMW-*ydcI*) was decreased compared with a vector-control strain (Δ*sucA*/pMW218) at the 1% significance level according to a *t*-test with the Bonferroni correction. By contrast, L-glutamic acid accumulation of the *ydcI* gene-deleted strain (Δ*sucA*Δ*ydcI*) was increased compared with the control strain (Δ*sucA*) at the 5% significance level using the same statistical test. These results suggest that YdcI represses *gltA* and *icdA* gene expression.

To confirm this regulation control, we measured enzyme activities of citrate synthase encoded by *gltA* in crude extracts of the four strains (Figure 8). The specific citrate synthase activity of the *ydcI* gene-amplified strain (Δ*sucA*/pMW-*ydcI*) was decreased compared with the vector control strain (Δ*sucA*/pMW218). The Bonferroni *t*-test revealed a 1% significant difference between the *ydcI* gene-amplified strain and a vector-control strain. By contrast, the citrate synthase activity of the *ydcI* gene-deleted strain (Δ*sucA*Δ*ydcI*) was increased compared with the control strain (Δ*sucA*). There was a 5% statistically significant difference between Δ*sucA*Δ*ydcI* and Δ*sucA*, suggesting that the decrease in L-glutamic acid accumulation by *ydcI* amplification was caused by a decrease in citrate synthase expression via repression of *gltA* by YdcI.

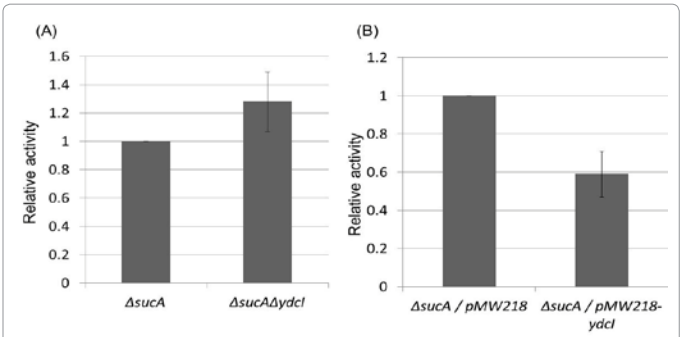

Figure 8: Activity of citrate synthase. The *t*-test with Bonferroni correction was performed. (A) Citrate synthase activity in *E. coli* MG1655 Δ*sucA*Δ*ydcI* relative to specific activity in MG1655 Δ*sucA*. The *p*-value was 0.027. (B) Citrate synthase activity in *E. coli* MG1655 Δ*sucA* pMW218-*ydcI* relative to specific activity in MG1655 Δ*sucA*/pMW218. The *p*-value was 0.008. Averages and standard deviations of three replicates are shown.

Discussion

Combining the identification of DNA motifs in the upstream regions of several genes with transcriptome data is a promising approach for biological network identification [8-10], and several algorithms and tools have been proposed to achieve this [11,32-34]. However, newly sequenced microbial genomes are usually annotated only for gene and RNA prediction. To overcome this, we hypothesized that there would be significant identity among nucleotide sequences of DNA motifs involved in the regulation of gene expression. Although the extraction of all possible DNA motifs with more than 50% identity was time consuming, it allowed us to predict a bilayer structure consisting of the ArcA regulon and YdcI regulon from DNA-array data.

Gene	*p*-value		
	0.0001	0.001	0.01
acnB	N	N	Y
acs	N	N	N
bglJ	N	N	N
fepA	N	N	Y
glcD	N	N	Y
gltA	Y	Y	Y
hdeA	Y	Y	Y
icd	Y	Y	Y
lpdA	Y	Y	Y
yojH	N	N	N
osmB	N	N	N
purE	N	N	Y
sdhC	Y	Y	Y
sucA	N	N	Y
ycgF	N	N	N
ydcI	N	N	N
yejG	N	N	N
yhhX	N	N	Y

Y, homologous sequence found; N, homologous sequence not found

Table 5: Identification of homologous region to DNA motif 20-150-1.

Strain	OD600	L-glutamic acid accumulation (g/L)	Residual glucose (g/L)
Δ*sucA*	9.8±0.5	14.9±1.0	0.5±0.9
Δ*sucA* / pMW218	9.7±0.5	16.2±1.0	0.2±0.3
Δ*sucA* / pMW-*ydcI*	10.0±0.5	13.8±0.5	0.4±0.5
Δ*sucA*Δ*ydcI*	10.3±0.4	16.1±0.7	0.1±0.1

Averages and standard deviations represent four replicates multiplied by three.

Table 6: L-glutamic acid fermentation analysis.

To test our scheme, we carried out ArcA-regulon analysis, particularly comparing known ArcA-binding motifs with predicted ArcA-binding sites. Recently, Tanaka et al. [35] evaluated the alignment and scoring of such a comparison. Here, we converted DNA motifs to a matrix with Shannon entropy [27], and used semi-global alignment with no penalty for end gaps to maximize the similarity score in pairwise alignment of DNA motifs [36]. This allowed us to assess similarity, and was useful for cluster analysis of DNA motifs. We concluded that our method was effective as we successfully extracted known ArcA-binding sites.

In *E. coli*, the expression of most genes that encode enzymes of the TCA cycle is regulated by ArcA. Thus, ArcA also regulates the carbon flux into the TCA cycle, which is why we recognize the ArcAB two-component system as an important factor for substance production. Here, we proposed that the putative transcription factor YdcI regulates *gltA*, *icd*, and *sdh* operon gene expression, and plays an important role in controlling the TCA cycle carbon flux. In low-GC-content Gram-positive bacteria, such as *Bacillus subtilis*, the global transcription factor CcpA and local transcription factor CcpC have been known to regulate genes encoding enzymes of the TCA cycle [37-39]. Lozada-Chávez et al. [40] analyzed the gene-regulatory network structure using a large-scale data set, and proposed that the gene-network hierarchy consists of global and local regulators, as seen in the relationships between *E. coli* ArcA and YdcI and *B. subtilis* CcpA and CcpC. Genome-sequence data revealed that *E. coli* ArcA was well-conserved and *E. coli* YdcI was partially conserved among γ-proteobacteria [41], whereas the *ydcI* gene was suggested to be acquired after divergence by gene duplication [40]. The TCA cycle local regulatory system in *E. coli* is suggested to have evolved in a biologically appropriate manner.

Both *E. coli* YdcI and *B. subtilis* CcpC are classified as LysR family transcriptional regulators, but do not share significant similarity at the amino-acid-sequence level. In general, DNA binding of such transcriptional regulatory proteins is affected by binding of low-molecular-weight compounds. Citrate was identified as an effector in

the case of CcpC in *B. subtilis* [39], and the identification of an effector for YdcI is a future goal.

The control of carbon flux into the TCA cycle is critical from the viewpoint of metabolic engineering in the field of substance production. Increases in carbon flux generate energy, and lead to an increase of biomass yield and *vice versa*. Theoretical carbon-flux analysis shows that a lower biomass yield will typically result in higher yield. For example, in L-lysine fermentation, a reduced carbon flux was expected to increase the L-lysine-production yield [42], and higher carbon flux into the TCA cycle is expected to lead to higher L-glutamic acid production [31]. In *E. coli*, *arcA* gene deletion increases the carbon flux into the TCA cycle [13], and we confirmed that this deletion (Δ*arcA*) indeed improves L-glutamic acid production in the *E. coli* MG1655 Δ*sucA* strain (unpublished data).

The current study suggested that YdcI repressed the expression of *gltA*, which encodes citrate synthase, and that Δ*ydcI* led to increased L-glutamic acid production. Extrapolating from this, *ydcI* gene amplification is likely to be a useful way to reduce the carbon flux into the TCA cycle. The repression of *gltA* gene expression by ArcA plays a major role in controlling carbon flow into the TCA cycle in the Δ*ydcI* strain. Indeed, L-glutamic acid production did not differ between the *arcA* and *ydcI* double mutant and the Δ*arcA* mutant (data not shown). However, because ArcA is a global regulator of more than 100 genes, global alteration in gene expression would not necessarily improve substance production in industrial strains because of the metabolism balance. Thus, differential modification of *arcA*and *ydcI* genes should be performed depending on the nature of the producing strains. Future studies will investigate YdcI-binding sites and YdcI effectors in order to characterize YdcI accurately. This will further our understanding of the control of carbon flux into the TCA cycle for improved application in the industrial fermentation process.

Acknowledgement

We thank Akira Imaizumi, Shintaro Iwatani, Yohei Yamada, and Takayuki Tanaka for invaluable discussions. We also thank Atsushi Matsuzawa and Yukiko Iwata for excellent technical support. This study was funded by Ajinomoto Co. Inc.

References

1. Galperin MY, Fernández-Suárez XM (2012) The 2012 Nucleic Acids Research Database Issue and the online Molecular Biology Database Collection. Nucleic Acids Res 40: D1-D8.
2. Pagani I, Liolios K, Jansson J, Chen IM, Smirnova T, et al. (2012) The Genomes OnLine Database (GOLD) v.4: status of genomic and metagenomic projects and their associated metadata. Nucleic Acids Res 40: D571-D579.
3. Stephanopoulos GN, Aristidou AA, Høiriis Nielsen J (1998)Metabolic engineering : principles and methodologies. Academic, London.
4. Iwatani S, Yamada Y, Usuda Y (2008) Metabolic flux analysis in biotechnology processes. Biotechnol Lett 30: 791-799.
5. Gelfand MS (2006) Evolution of transcriptional regulatory networks in microbial genomes. Curr Opin Struct Biol 16: 420-429.
6. Barrett CL, Cho BK, Palsson BO (2011) Sensitive and accurate identification of protein-DNA binding events in ChIP-chip assays using higher order derivative analysis. Nucleic Acids Res 39: 1656-1665.
7. Ishihama A (2010) Prokaryotic genome regulation: multifactor promoters, multitarget regulators and hierarchic networks. FEMS Microbiol Rev 34: 628-645.
8. D'haeseleer P (2006) How does DNA sequence motif discovery work? Nat Biotechnol 24: 959-961.
9. Masuda N, Church GM (2003) Regulatory network of acid resistance genes in Escherichia coli. Mol Microbiol 48: 699-712.
10. Zhu C, Byers KJ, McCord RP, Shi Z, Berger MF, et al. (2009) High-resolution DNA-binding specificity analysis of yeast transcription factors. Genome Res 19: 556-566.
11. Tompa M, Li N, Bailey TL, Church GM, De Moor B, et al. (2005) Assessing computational tools for the discovery of transcription factor binding sites. Nat Biotechnol 23: 137-144.
12. Salmon KA, Hung SP, Steffen NR, Krupp R, Baldi P, et al. (2005) Global gene expression profiling in Escherichia coli K12: effects of oxygen availability and ArcA. J Biol Chem 280: 15084-15096.
13. Waegeman H, Beauprez J, Moens H, Maertens J, De Mey M, et al. (2011) Effect of iclR and arcA knockouts on biomass formation and metabolic fluxes in Escherichia coli K12 and its implications on understanding the metabolism of Escherichia coli BL21 (DE3). BMC Microbiol 11: 70.
14. Oshima T, Aiba H, Masuda Y, Kanaya S, Sugiura M, et al. (2002) Transcriptome analysis of all two-component regulatory system mutants of Escherichia coli K-12. Mol Microbiol 46: 281-291.
15. Imaizumi A, Kojima H, Matsui K (2006) The effect of intracellular ppGpp levels on glutamate and lysine overproduction in Escherichia coli. J Biotechnol 125: 328-337.
16. Datsenko KA, Wanner BL (2000) One-step inactivation of chromosomal genes in Escherichia coli K-12 using PCR products. Proc Natl Acad Sci USA 97: 6640-6645.
17. Katashkina ZhI, Skorokhodova AIu, Zimenkov DV, Gulevich AIu, Minaeva NI, et al. (2005) [Tuning of expression level of the genes of interest located in the bacterial chromosome]. Mol Biol (Mosk) 39: 823-831.
18. Doroshenko V, Airich L, Vitushkina M, Kolokolova A, Livshits V, et al. (2007) YddG from Escherichia coli promotes export of aromatic amino acids. FEMS Microbiol Lett 275: 312-318.
19. Weitzman PDJ (1969) Citrate synthase from *Escherichia coli* [EC 4.1.3.7 Citrate oxaloacetage-lyase (CoA-acetylating)]. In Lowenstein JM (edn), Methods in Enzymology, Elsevier 13: 22-26.
20. de Hoon MJL, Imoto S, Nolan J, Miyano S (2004) Open source clustering software. Bioinformatics 20: 1453-1454.
21. Bailey TL, Williams N, Misleh C, Li WW (2006) MEME: discovering and analyzing DNA and protein sequence motifs. Nucleic Acids Res 34: W369-W373.
22. Crooks GE, Hon G, Chandonia JM, Brenner SE (2004) WebLogo: a sequence logo generator. Genome Res 14: 1188-1190.
23. Keseler IM, Collado-Vides J, Gama-Castro S, Ingraham J, Paley S, et al. (2005) EcoCyc: a comprehensive database resource for Escherichia coli. Nucleic Acids Res 33: D334-D337.
24. McCue LA, Thompson W, Carmack CS, Lawrence CE (2002) Factors influencing the identification of transcription factor binding sites by cross-species comparison. Genome Res12: 1523-1532.
25. Robison K, McGuire AM, Church GM (1998) A comprehensive library of DNA-binding site matrices for 55 proteins applied to the complete Escherichia coli K-12 genome. J Mol Biol 284: 241-254.
26. Salgado H, Gama-Castro S, Martínez-Antonio A, Díaz-Peredo E, Sánchez-Solano F, et al. (2004) RegulonDB (version 4.0): transcriptional regulation, operon organization and growth conditions in Escherichia coli K-12. Nucleic Acids Res 32: D303-D306.
27. Schneider TD, Stormo GD, Gold L, Ehrenfeucht A (1986) Information content of binding sites on nucleotide sequences. J Mol Biol 188: 415-431.
28. Smith TF, Waterman MS (1981) Identification of common molecular subsequences. J Mol Biol 147: 195-197.
29. McGuire AM, De Wulf P, Church GM, Lin EC (1999) A weight matrix for binding recognition by the redox-response regulator ArcA-P of Escherichia coli. Mol Microbiol 32: 219-221.
30. Chao G, Shen J, Tseng CP, Park SJ, Gunsalus RP (1997) Aerobic regulation of isocitrate dehydrogenase gene (icd) expression in Escherichia coli by the arcA and fnr gene products. J Bacteriol 179: 4299-4304.
31. Usuda Y, Nishio Y, Iwatani S, Van Dien SJ, Imaizumi A, et al. (2010) Dynamic modeling of Escherichia coli metabolic and regulatory systems for amino-acid production. J Biotechnol 147: 17-30.
32. Blom EJ, Roerdink JB, Kuipers OP, van Hijum SA (2009) MOTIFATOR: detection

and characterization of regulatory motifs using prokaryote transcriptome data. Bioinformatics 25: 550-551.

33. Halperin Y, Linhart C, Ulitsky I, Shamir R (2009) Allegro: analyzing expression and sequence in concert to discover regulatory programs. Nucleic Acids Res 37: 1566-1579.

34. Mrázek J (2009) Finding sequence motifs in prokaryotic genomes--a brief practical guide for a microbiologist. Brief Bioinform 10: 525-536.

35. Tanaka E, Bailey T, Grant CE, Noble WS, Keich U (2011) Improved similarity scores for comparing motifs. Bioinformatics 27: 1603-1609.

36. Setubal J, Meidanis J (1997) Introduction to computational molecular biology. PWS Publishing Company, Boston, MA, USA.

37. Fujita Y (2009) Carbon catabolite control of the metabolic network in Bacillus subtilis. Biosci Biotechnol Biochem 73: 245-259.

38. Kim HJ, Roux A, Sonenshein AL (2002) Direct and indirect roles of CcpA in regulation of Bacillus subtilis Krebs cycle genes. Mol Microbiol 45: 179-190.

39. Kim SI, Jourlin-Castelli C, Wellington SR, Sonenshein AL (2003) Mechanism of repression by Bacillus subtilis CcpC, a LysR family regulator. J Mol Biol 334: 609-624.

40. Lozada-Chávez I, Angarica VE, Collado-Vides J, Contreras-Moreira B (2008) The role of DNA-binding specificity in the evolution of bacterial regulatory networks. J Mol Biol379: 627-643.

41. Jennings ME, Quick LN, Soni A, Davis RR, Crosby K, et al. (2011) Characterization of the Salmonella enterica serovar Typhimurium ydcI gene, which encodes a conserved DNA binding protein required for full acid stress resistance. J Bacteriol 193: 2208-2217.

42. Kiss RD, Stephanopoulos G (1992) Metabolic characterization of a L-lysine-producing strain by continuous culture. Biotechnol Bioeng 39: 565-574.