Conservation of Gene Co-Regulation between Two Prokaryotes: *Bacillus subtilis* and *Escherichia coli*

Shujiro Okuda[1]
okuda@kuicr.kyoto-u.ac.jp

Shuichi Kawashima[2]
shuichi@hgc.jp

Susumu Goto[1]
goto@kuicr.kyoto-u.ac.jp

Minoru Kanehisa[1]
kanehisa@kuicr.kyoto-u.ac.jp

[1] Bioinformatics Center, Institute for Chemical Research, Kyoto University, Gokasho, Uji, Kyoto 611-0011, Japan
[2] Human Genome Center, Institute of Medical Science, University of Tokyo, 4-6-1 Shirokane-dai, Minato-ku, Tokyo 108-8639, Japan

Abstract

We measured conservation of gene co-regulation between two distantly related prokaryotes, *B. subtilis* and *E. coli.* The co-regulation between genes was extracted from knowledge of regulation of genes stored in databases. For *B. subtilis* operons, we obtained the data set from ODB which we have developed and, for the regulons, we used DBTBS. For *E. coli* data set, we used known regulons derived from RegulonDB. We obtained a reliable data set of co-regulated genes in *B. subtilis* and *E. coli.* About 60-80 % of gene pairs conserved co-regulation relationships, so co-regulation between genes are highly conserved even between distantly related species. To measure the functional relationship between these conserved genes, we used KEGG PATHWAY and COG. When two co-regulated genes are in the same biological pathway in KEGG or share the same functional category in COG, we assume that they have the same function. As a result, we also found that many conserved co-regulated gene pairs share the same functions. These observations would help to predict gene co-regulation and protein functions.

Keywords: operon, regulon, gene co-regulation, *Bacillus subtilis*, *Escherichia coli*, database

1 Introduction

The increasing availability of complete genomes enables us to perform large scale comparative genomics. It has been demonstrated that functionally related genes are often clustered on the genome [8, 15, 24]. These gene clusters are often co-transcribed as operons, which are defined as a series of genes encoded in the same strand and transcribed into one mRNA. Operon structures are basically known to be one of the transcriptional regulatory mechanisms in prokaryotes, but a similar mechanism was also reported in a few eukaryotes [2, 12]. The genes transcribed in an operon are functionally related and the proteins are part of the same protein complex or metabolic pathway [14]. In addition, some operons are often co-regulated by the same transcriptional factor as a regulon. The mechanism of regulons also plays an important role in the function of the organism. Understanding such an intricate gene transcriptional system in prokaryotes should enhance our knowledge of the function and regulation of genes.

To understand genome organization, information regarding the co-regulation of genes derived from the literature and experimental data have been accumulated in some databases. For example, RegulonDB has been developed to store information about the regulation of genes in *Escherichia coli* [20]. For *Bacillus subtilis*, DBTBS has collected transcription factors and regulated genes [13]. We have developed a database of operons, named ODB (`http://odb.kuicr.kyoto-u.ac.jp`), which currently includes the known information of operons derived from the literature and experiments.

In addition to the accumulated knowledge about the co-regulation of genes, various computational prediction methods of operons and regulons have been developed [3, 4, 5, 7, 8, 17, 18, 19, 23, 28]. Comparative genomics approaches have claimed that the co-regulation of genes as operons are likely to collapse in evolutionary history [6, 9, 27]. The instability of operon structures imply that the co-regulation of genes may not be conserved. Thus, Teichmann *et al.* have shown that the co-regulation of genes are not likely to be conserved according to the analysis of gene neighborhood [26]. However, this issue is in controversy. Snel *et al.* claimed that even if genes in an operon in a species are not in a operon in another species, they can still be co-regulated as a regulon. They have shown that the co-regulation of genes are highly conserved between prokaryotes and eukaryotes [22].

To identify conserved co-regulated genes between the completely sequenced Gram-positive and Gram-negative bacteria, *B. subtilis* [11] and *E. coli* [1], we used the data set of known operons and regulons based on documented information obtained from the databases of DBTBS, ODB and RegulonDB and estimated the conservation of co-regulation between these distantly related species. In this study, we report that co-regulated genes are likely to be conserved between distantly related species.

2 Methods

2.1 Genomes of *B. subtilis* and *E. coli*

We obtained the *B. subtilis* and *E. coli* genome data from KEGG [10]. Genes from these genomes are linked to biological pathways in KEGG PATHWAY. To identify orthologs between these genes, we used SSDB, which stores the results of homology searches by the Smith-Waterman algorithm [16, 21]. We regarded the bi-directional best hit (BBH) as the ortholog gene between two organisms.

2.2 Known Operons and Regulons

We have been developing the operon database, named ODB (`http://odb.kuicr.kyoto-u.ac.jp`), which stores about 700 known operons derived from the literature and experimental data. The current version of this database has the data set of *B. subtilis* operons. For the regulons of *B. subtilis*, we use the data set obtained from DBTBS [13]. This database contains the known transcription factors and regulated genes of *B. subtilis*. Combining these two data sets, we obtained the set of co-regulated genes in *B. subtilis*. For regulons and operons in *E. coli*, we obtained the known regulons from RegulonDB [20].

2.3 Known Co-Regulated Gene Pairs

We defined two types of co-regulated gene pairs which we call 'Close Pair (CP)' and 'Distant Pair (DP)' (Figure 1). CP is a gene pair in which both genes are in the same operon. DP is one in which they are in different operons but belong to the same regulon. We assume that the operons and regulons obtained from the literature and wet experiments are co-regulated. In contrast to microarray experimental data containing numerical values indicating varying degrees of co-regulation, we used known sets of co-regulated genes from these databases as a reliable data set.

3 Results and Discussion

3.1 Determining Gene Co-Regulation in *B. subtilis* and *E. coli*

Co-regulated genes have successfully been determined by a large number of known operons and regulons (Table 1). Known operons of *B. subtilis* were 689 and *E. coli* were 683. These operons consist of 1163 genes in *B. subtilis* and 1467 genes in *E. coli*. Genes in an operon are co-transcribed as single

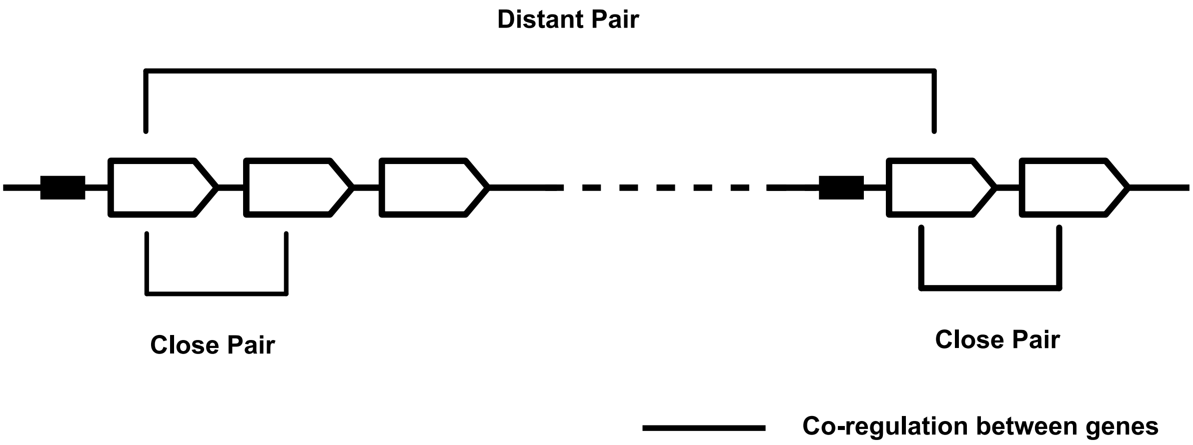

Figure 1: Illustration of two types of gene co-regulation. Gene pairs in the same operon are defined as 'Close Pairs (CPs)'. Gene pairs in the same regulon but in different operons are defined as 'Distant Pairs (DPs)'. The three open boxes on the left and the two on the right indicate operons, and they are regulated by the same transcription factor.

mRNA, so all the gene pairs in an operon are always co-regulated without a few exceptions such as in the case where the transcription is intricately regulated by multiple regulatory elements such as trancription factors and internal terminators. In this study, we defined all gene pairs in a known operon as co-regulated gene pairs. These co-regulated gene pairs are 'closely' located on the genome, so we call them Close Pair (CP). Thus, we obtained 1932 CPs of *B. subtilis* and 2349 CPs of *E. coli.*

On the other hand, regulons are the mechanism for co-regulation of distantly located gene pairs on a genome. We obtained 97 known regulons of *B. subtilis* and 129 of *E. coli.* For a regulon, gene pairs across operons belonging to it are co-regulated. These gene pairs are not close on the genome because they are not in the same operon, so we call them Distant Pairs (DPs). Thus, we obtained 121921 DPs in *B. subtilis* and 39905 DPs in *E. coli.*

Table 1: Statistics of co-regulated genes in *B. subtilis* and *E. coli.*

	B. subtilis	*E. coli*
Regulon	97	129
Operon	689	683
Gene	1163	1467
Close Pair	1932	2349
Distant Pair	121921	39905

3.2 Conservation of Co-Regulated Gene Pair

We explored the conservation of co-regulation between genes in *B. subtilis* and *E. coli.* When a gene pair is co-regulated in a species and the ortholog genes are co-regulated in another species, the gene co-regulation is considered to be conserved. We summarize the conservation of Close Pairs and Distant Pairs in *B. subtilis* and *E. coli* in Table 2 and illustrate the relationships between CPs and DPs in Figure 2. Of the 1932 CPs in *B. subtilis*, 553 gene pairs were detected as ortholog pairs in *E. coli.* Of these 553 gene pairs, 412 gene pairs were in known operons in *E. coli.* The regulation of these 412 gene pairs in *E. coli* have been described in RegulonDB. Of these 412 gene pairs, 260 (202+58) gene

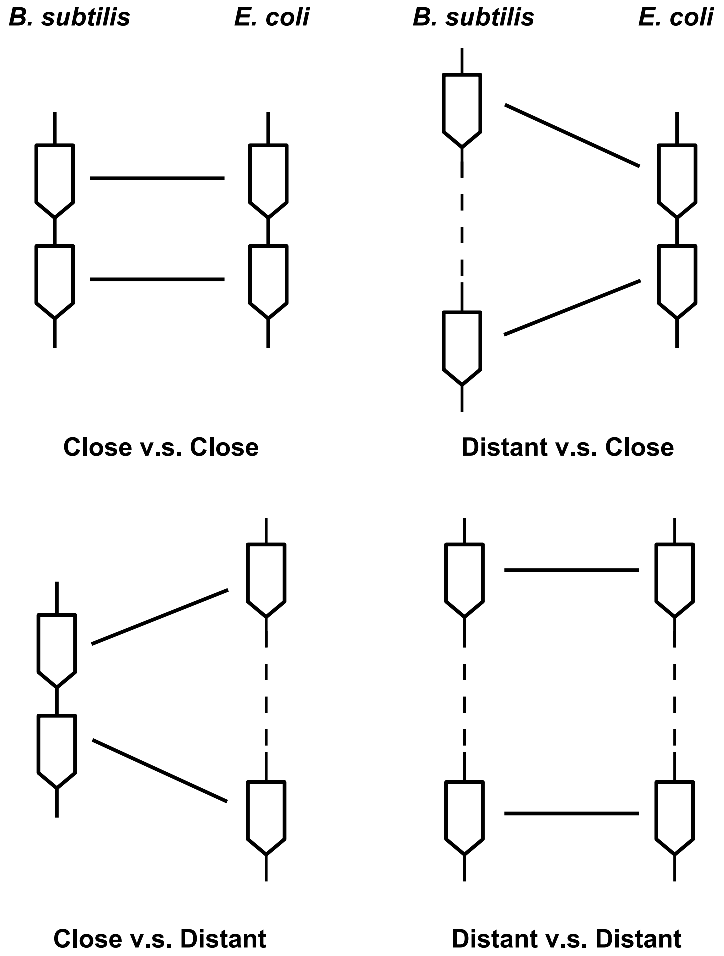

Figure 2: Illustration of four types of relationships of gene co-regulation. We classified gene co-regulation into four types with the criterion of co-regulation as gene pairs in a operon (CP) and co-regulation as gene pairs in the same regulon but in different operons (DP). 'Close vs. Close' indicates that a CP in *B. subtilis* is co-regulated as a CP in *E. coli*. 'Close vs. Distant' and 'Distant vs. Close' indicate that a CP in a species is co-regulated as a DP in another species. 'Distant vs. Distant' indicates that a DP in *B. subtilis* is co-regulated as a DP in *E. coli*.

pairs were in the same *E. coli* regulons. Thus, about 60% of the gene pairs that are in an operon in *B. subtilis* are also co-regulated in *E. coli.* Of these conserved co-regulated gene pairs, 202 were CPs and 58 were DPs. So, co-regulation between genes were more frequently conserved in operons than in regulons.

Table 2: Co-regulated gene pairs conserved between *B. subtilis* and *E. coli.*

B. subtilis vs. *E. coli*	Number of gene pairs	Number of genes
Close vs. Close	202	136
Close vs. Distant	58	30
Distant vs. Close	15	25
Distant vs. Distant	395	110

On the other hand, of 2349 CPs in *E. coli*, 624 ortholog gene pairs in *B. subtilis* were detected. Of these 624 gene pairs in *B. subtilis*, 267 gene pairs were in known regulons; they were detected in ODB and DBTBS. Of these 267 gene pairs, 217 (202+15) gene pairs were in the same regulons and operons. Thus, about 80% of the gene pairs that are in the same operon in *E. coli* had conserved co-regulation. Of these 217 gene pairs, 202 were CPs and 15 were DPs in *E. coli.* So, co-regulation was less conserved in CP-DP relationships.

Distant Pairs in two prokaryotes are also gene pairs co-regulated as well as CPs but distantly located on a genome. We tested the same analysis to estimate the conservation of co-regulation between genes in DPs. Of these 121921 DPs in *B. subtilis*, 22320 gene pairs had ortholog pairs in *E. coli.* Of these 22320 gene pairs, 9584 were in the data set of regulons in *E. coli.* Analysis of these 9584 genes showed that 410 (395+15) gene pairs were in the same regulons in *E. coli.* We tested the DPs in *E. coli.* The 3512 gene pairs in *B. subtilis* of the 39905 DPs in *E. coli* were ortholog pairs and 1480 gene pairs were in the same regulons in *B. subtilis.* Of these1480 gene pairs, 453 (395+58) were in the same regulons. Thus, distant co-regulation in a regulon in a species was less conserved in another distantly related species. This estimate, however, may be an underestimation for Distant Pairs. The estimate method used here is too sensitive to loss and gain of genes co-regulated in a regulon, because the gain of an gene in a regulon causes a great increase in the number of DPs in the regulon and vice versa.

Here, we used the same method reported by Snel *et al.* [22] to estimate the conservation of gene co-regulation between two prokaryotes. They estimated that 80% of the gene pairs in the same operon in *B. subtilis* are co-regulated in *E. coli*, based on known operons in *B. subtilis* and known regulons in *E. coli*, and they also mentioned that the gene co-regulation is highly conserved. Furthermore, they claimed that half of the conserved co-regulated genes can be linked in so-called ‘runs’ [15] and taking regulon data into account leads to about a two-fold increase in conservation [22]. However, these observations do not distinguish between the co-regulation in the same operons and the co-regulation between distant genes in the same regulon. Thus, in our work, we classified co-regulated gene pairs into CPs and DPs to make these differences clear. Moreover, our experiments used not only known operon data in *B. subtilis* but also additional regulon data that they did not use. While we took regulon data into account, the conservation as operons dominated between these two distantly related species. We did find that gene pairs co-regulated in a operon that were distantly located on another genome, and which were conserved as a regulon. Furthermore, for DPs in a species, the co-regulation could often be conserved as a DP in another species. This implies that regulon structure is also conserved between two distantly related species. However, of all the DPs we obtained, these conserved DPs were actually rather few. Although some examples of increase in conservation between DP pairs could be found, in general this fraction was much smaller. This suggests that the combination of operons in a regulon is more likely to change, so only a part of them are conserved.

3.3 Conservation of Co-Regulation of Genes in Terms of Orthologs

We summarize the number of ortholog genes whose co-regulation is conserved between these two species in Table 2. For example, when a CP in *B. subtilis* was conserved as a CP in *E. coli* (Close vs. Close), the number of ortholog genes was 136. The DP gene pairs in a species conserved as a CP in another species numbered a small fraction, i.e. 25 and 30. However, the number of DP genes in a species that were conserved as a DP in another species numbered 110. There were 208 conserved co-regulated genes in total (Table 3). 1255 ortholog genes between *B. subtilis* and *E. coli* were obtained as BBH (Table 3). Of these 1255 orthologs, there were 292 gene pairs in which both genes were detected in known operons or regulons. Thus, when a gene pair is co-regulated, about 70% (208 of 292) of their ortholog genes in another species conserved the co-regulation with at least one counterpart gene in co-regulation relationships even between distantly related species. A gene in an operon is multiply counted as co-regulated gene pairs as the number of genes in the operon increase. This overcount is true in regulons. However, it is interesting that about 70% of orthologs conserved at least one of the co-regulation relationships and about 60-80% of gene pairs that are in the same operon are also co-regulated in distantly related species.

Table 3: Conservation of bi-directional best hits using Smith-Waterman score between *B. subtilis* and *E. coli.*

	Number of bi-directional best hits
Co-regulation conserved	208
Both genes assigned to known data	292
One genes assigned to known data	793
All bi-directional best hits	1255

3.4 Functions of Conserved Co-Regulated Genes

To measure the functional relationships between the conserved co-regulated gene pairs between *B. subtilis* and *E. coli*, we used KEGG PATHWAY, which store knowledge on molecular interaction networks in biological processes [10], and COG, which includes ortholog protein groups with functional families [25]. When gene pairs map to the same biological pathway or share the same COG functional categories, we assume that they are functionally related. With this criterion, about 60 % of the conserved gene pairs were found to be functionally related (395 / 664 in *B. subtilis* and 397 / 659 in *E. coli*). The number of conserved genes in the KEGG biological pathways and the fraction among all conserved genes is shown in Table 4. The sum of these fraction is greater than 1 because some genes are assigned to multiple biological pathways. Conserved co-regulated genes were observed in various biological pathways, but they seem to appear most in the pathways essential to cell viability.

4 Conclusion

We have developed a database of operons, named ODB, which includes the known operon information derived from literature and experimental data. Using the known operons from this database and the known regulons from DBTBS and RegulonDB, we could successfully detect the co-regulated genes between *B. subtilis* and *E. coli.* We analyzed the co-regulation relationships between the conserved genes across distantly related species from known information of operons and regulons on a large scale. To measure the conservation of gene co-regulation, we defined the co-regulated gene pairs as CPs which are co-regulated in a operon and DPs which are co-regulated across operons. As a result, 60-80% of co-regulated gene pairs were conserved between these two species and this estimate supports

Table 4: Conserved co-regulated gene pairs in biological pathways.

Pathway	Number of genes	Fraction of total
Carbohydrate Metabolism	68	0.33
Amino Acid Metabolism	64	0.31
Energy Metabolism	56	0.27
Nucleotide Metabolism	46	0.22
Membrane Transport	41	0.20
Metabolism of Cofactors and Vitamins	25	0.12
Translation	23	0.11
Signal Transduction	18	0.09
Biosynthesis of Polyketides and Nonribosomal Peptides	14	0.07
Others	33	0.16

the previous work [22], but CP pairs were more likely to be conserved as CPs than DPs in another species and this estimate was much larger than previous estimate [22]. However, in terms of the number of ortholog genes, we also found that about 70% of co-regulated ortholog genes are conserved between these distantly related species. From analyses of biological pathways and functional categories of genes, we found that conserved co-regulated gene pairs tend to share the same functions.

Acknowledgments

We thank Kiyoko F. Aoki-Kinoshita for critical reading of our manuscript. This work was supported by grants from the Ministry of Education, Culture, Sports, Science and Technology, the Japan Society for the Promotion of Science, and the Japan Science and Technology Corporation. The computational resource was provided by the Bioinformatics Center, Institute for Chemical Research, Kyoto University and the Super Computer System, Human Genome Center, Institute of Medical Science, University of Tokyo.

References

[1] Blattner, F. R., Plunkett, G. 3rd., Bloch, C. A., Perna, N. T., Burland, V., Riley, M., Collado-Vides, J., Glasner, J. D., Rode, C. K., Mayhew, G. F., Gregor, J., Davis, N. W., Kirkpatrick, H. A., Goeden, M. A., Rose, D. J., Mau, B., and Shao, Y., The complete genome sequence of *Escherichia coli* K-12, *Science*, 277(5331):1453–1474, 1997.

[2] Blumenthal, T., Evans, D., Link, C. D., Guffanti, A., Lawson, D., Thierry-Mieg, J., Thierry-Mieg, D., Chiu, W. L., Duke, K., Kiraly, M., and Kim, S. K., A global analysis of *Caenorhabditis elegans* operons, *Nature*, 417(6891):851–854, 2002.

[3] Bockhorst, J., Craven, M., Page, D., Shavlik, J., and Glasner, J., A bayesian network approach to operon prediction, *Bioinformatics*, 19(10):1227–1235, 2003.

[4] Bockhorst, J., Qiu, Y., Glasner, J., Liu, M., Blattner, F., and Craven, M., Predicting bacterial transcription units using sequence and expression data, *Bioinformatics*, 19 Suppl 1:i34–43, 2003.

[5] Craven, M., Page, D., Shavlik, J., Bockhorst, J., and Glasner, J., A probabilistic learning approach to whole-genome operon prediction, *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, 8:116–127, 2000.

[6] Dandekar, T. Snel, B., Huynen, M., and Bork, P., Conservation of gene order: a fingerprint of proteins that physically interact, *Trends Biochem. Sci.*, 23:324–328, 1998.

[7] De Hoon, M. J., Imoto, S., Kobayashi, K., Ogasawara, N., and Miyano, S., Predicting the operon structure of *Bacillus subtilis* using operon length, intergene distance, and gene expression information, *Pac. Symp. Biocomput.*, 9:276–287, 2004.

[8] Ermolaeva, M. D., White, O., and Salzberg, S. L., Prediction of operons in microbial genomes, *Nucleic Acids Res.*, 29(5):1216–1221, 2001.

[9] Itoh, T., Takemoto, K., Mori, H., and Gojobori, T., Evolutionary instability of operon structures disclosed by sequence comparisons of complete microbial genomes, *Mol. Biol. Evol.*, 16(3):332–346, 1999.

[10] Kanehisa, M., Goto, S., Kawashima, S., Okuno, Y., and Hattori, M., The KEGG resource for deciphering the genome, *Nucleic Acids Res.*, 32 Database issue:D277–280, 2004.

[11] Kunst, F., Ogasawara, N., Moszer, I., Albertini, A. M., Alloni, G., Azevedo, V. Bertero, M. G., Bessieres, P., Bolotin, A., Borchert, S., Borriss, R., Boursier, L., Brans, A., Braun, M., Brignell, S. C., Bron, S., Brouillet, S., Bruschi, C. V., Caldwell, B., Capuano, V., Carter, N. M., Choi, S. K., Codani, J. J., Connerton, I. F., Danchin, A., and *et al.*, The complete genome sequence of the gram-positive bacterium *Bacillus subtilis*, *Nature*, 390(6657):249–256, 1997.

[12] Lercher, M. J., Blumenthal, T., and Hurst, L. D., Coexpression of neighboring genes in *Caenorhabditis elegans* is mostly due to operons and duplicate genes, *Genome Res.*, 13(2):238–243, 2003.

[13] Makita, Y., Nakao, M., Ogasawara, N., and Nakai, K., DBTBS: database of transcriptional regulation in *Bacillus subtilis* and its contribution to comparative genomics, *Nucleic Acids Res.*, 32 Database issue:D75–77, 2004.

[14] Ogata, H., Fujibuchi, W., Goto, S., and Kanehisa, M., A heuristic graph comparison algorithm and its application to detect functionally related enzyme clusters, *Nucleic Acids Res.*, 28(20):4021–4028, 2000.

[15] Overbeek, R., Fonstein, M., D'Souza, M., Pusch, G. D., and Maltsev, N., The use of gene clusters to infer functional coupling, *Proc. Natl. Acad. Sci. USA*, 96(6):2896–2901, 1999.

[16] Pearson, W. R., Searching protein sequence libraries: comparison of the sensitivity and selectivity of the Smith-Waterman and FASTA algorithms, *Genomics*, 11:635–650, 1991.

[17] Price, M. N., Huang, K. H., Alm, E. J., and Arkin, A. P., A novel method for accurate operon predictions in all sequenced prokaryotes, *Nucleic Acids Res.*, 33(3):880–892, 2005.

[18] Sabatti, C., Rohlin, L., Oh, M. K., and Liao, J. C., Co-expression pattern from DNA microarray experiments as a tool for operon prediction, *Nucleic Acids Res.*, 30(13):2886–2893, 2002.

[19] Salgado, H., Moreno-Hagelsieb, G., Smith, T. F., and Collado-Vides, J., Operons in *Escherichia coli*: genomic analyses and predictions, *Proc. Natl. Acad. Sci. USA*, 97(12):6652–6657, 2000.

[20] Salgado, H., Santos-Zavaleta, A., Gama-Castro, S., Millan-Zarate, D., Diaz-Peredo, E., Sanchez-Solano, F., Perez-Rueda, E., Bonavides-Martinez, C., and Collado-Vides, J., RegulonDB (version 3.2): transcriptional regulation and operon organization in *Escherichia coli* K-12, *Nucleic Acids Res.*, 29(1):72–74, 2001.

[21] Smith, T. F. and Waterman, M. S., Identification of common molecular subsequences, *J. Mol. Biol.*, 147:195–197, 1981.

[22] Snel, B., van Noort, V., and Huynen, M. A., Gene co-regulation is highly conserved in the evolution of eukaryotes and prokaryotes, *Nucleic Acids Res.*, 32(16):4725–4731, 2004.

[23] Steinhauser, D., Junker, B. H., Luedemann, A., Selbig, J., and Kopka, J., Hypothesis-driven approach to predict transcriptional units from gene expression data, *Bioinformatics*, 20(12):1928–1939, 2004.

[24] Tamames, J., Casari, G., Ouzounis, C., and Valencia, A., Conserved clusters of functionally related genes in two bacterial genomes, *J. Mol. Evol.*, 44(1):66–73, 1997.

[25] Tatusov, R. L., Koonin, E. V., and Lipman, D. J., A genomic perspective on protein families, *Science*, 278(5338):631–637, 1997.

[26] Teichmann, S. A. and Babu, M. M., Conservation of gene co-regulation in prokaryotes and eukaryotes, *Trends Biotechnol.*, 20(10):407–410, 2002.

[27] Wolf, Y. I., Rogozin, I. B., Kondrashov, A. S., and Koonin, E. V., Genome alignment, evolution of prokaryotic genome organization, and prediction of gene function using genomic context, *Genome Res.*, 11(3):356–372, 2001.

[28] Yada, T., Nakao, M., Totoki, Y., and Nakai, K., Modeling and predicting transcriptional units of *Escherichia coli* genes using hidden markov models, *Bioinformatics*, 15(12):987–993, 1999.